PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

FÁBIO MOREIRA FREITAS DA SILVA

# ENTIDADES NOMEADAS E EXTRAÇÃO DE INFORMAÇÃO NO AUXÍLIO ÀS INVESTIGAÇÕES DE CRIMES DE LAVAGEM DE DINHEIRO

Porto Alegre
2020

PÓS-GRADUAÇÃO - *STRICTO SENSU*


Pontifícia Universidade Católica
do Rio Grande do Sul

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL**
**ESCOLA POLITÉCNICA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

# ENTIDADES NOMEADAS E EXTRAÇÃO DE INFORMAÇÃO NO AUXÍLIO ÀS INVESTIGAÇÕES DE CRIMES DE LAVAGEM DE DINHEIRO

**FÁBIO MOREIRA FREITAS DA SILVA**

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profa. Dra. Renata Vieira

**Porto Alegre**
**2020**

# Ficha Catalográfica

S586e Silva, Fábio Moreira Freitas da

Entidades nomeadas e extração de informação no auxílio às investigações de crimes de lavagem de dinheiro / Fábio Moreira Freitas da Silva . – 2020.

70 p.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Renata Vieira.

1. Processamento de Linguagem Natural. 2. Crime. 3. Lavagem de dinheiro. 4. Investigação. 5. Polícia Federal. I. Vieira, Renata. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

Fábio Moreira Freitas da Silva

**Entidades nomeadas e extração de informação no auxílio às investigações de crimes de lavagem de dinheiro**

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado em 12 de março de 2020.

**BANCA EXAMINADORA:**

Profa. Dra. Daniela Barreiro Claro (UFBA)

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS)

Profa. Dra. Renata Vieira (PPGCC/PUCRS - Orientadora)

## DEDICATÓRIA

Dedico este trabalho primeiramente a Deus, por ser essencial em minha vida.

À minha família que, com muito amor e compreensão, me apoiou e me deu forças para que eu chegasse até esta etapa de minha vida.

Aos professores Dra. Renata Vieira, Dr. Rafael Heitor Bordini e Dr. Avelino Francisco Zorzo, seres humanos excepcionais, que me apoiaram e me mostraram o caminho para o conhecimento.

Aos meus colegas de trabalho, Alexandre da Silveira Isbarrola, Alessandro Maciel Lopes, Leandro Dias Cunha, Rafael Scorsatto Ortiz e Adelson Cabral de Sena, pelo apoio e compreensão.

Aos meus colegas Olimar Teixeira Borges, Joaquim Francisco dos Santos Neto e Henrique Dias Pereira dos Santos, e à todas as pessoas que de alguma forma acreditaram e me apoiaram, tornando possível transformar um sonho em realidade.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal Nível Superior – Brasil (CAPES), Código de Financiamento 001, e do Instituto Nacional de Ciência e Tecnologia Forense - INCT Forense.

“Nós somos o que repetidamente fazemos. A excelência, então, não é um ato, mas um hábito.”
(Aristóteles)

# ENTIDADES NOMEADAS E EXTRAÇÃO DE INFORMAÇÃO NO AUXÍLIO ÀS INVESTIGAÇÕES DE CRIMES DE LAVAGEM DE DINHEIRO


## RESUMO

Com o crescente avanço tecnológico, as organizações criminosas estão cada vez mais utilizando a tecnologia para o cometimento de crimes, especialmente o crime de lavagem de dinheiro, previsto na Lei nº 9.613/98 e alterado pela Lei nº 12.683/12, em razão de sua complexidade e sofisticação. O volume de informações, provenientes de fontes abertas e materiais apreendidos pelos órgãos de segurança pública, em especial pela Polícia Federal, apresenta um desafio para a análise investigativa. Visando oferecer maior suporte tecnológico às investigações policiais, este trabalho apresenta um estudo aplicado de modelos de Reconhecimento de Entidades Nomeadas (REN), que consiste em localizar e categorizar nomes importantes e nomes próprios em textos livres, e de outras técnicas de Processamento de Linguagem Natural (PLN) e de visualização de informações. Foram realizados experimentos a partir de peças policiais produzidas e fornecidas pela Polícia Federal. Os resultados demonstram possibilidades para aplicação das técnicas computacionais na identificação de elementos que indiquem a autoria e a materialidade criminosa, auxiliando as equipes de investigação na elucidação de crimes complexos, como é o caso do crime de lavagem de dinheiro.



**Palavras-Chave:** Processamento de Linguagem Natural, crime, lavagem de dinheiro, investigação, Polícia Federal.

# NAMED ENTITIES AND INFORMATION EXTRACTION TO ASSIST INVESTIGATIONS OF MONEY LAUNDERING CRIMES


## ABSTRACT

With the increasing technological advancement, as criminal organizations are increasingly using technology to commit crimes, especially the crime of money laundering, provided for in Law 9613/98 and amended by Law 12683/12, due to its complexity and sophistication. The volume of information, coming from open sources and materials seized by public security bodies, especially by the Federal Police, presents a challenge for an investigative analysis. In order to offer greater technological support to investigations of agents, this work presents a study applied to the Named Entity Recognition (NER) models, which consists of locating and categorizing important names and unique names in free texts and other Natural Language Processing techniques (NLP) and information visualization. Experiments were carried out using parts produced and supplied by the Federal Police. The results demonstrate the possibilities of applying computational techniques in the identification of elements that include a criminal authority and material, assist as investigation teams in the elucidation of complex crimes, such as money laundering.



**Keywords:** Natural Language Processing, crime, money laundering, investigation, Federal Police.

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE SIGLAS

AQI – Auto de Qualificação e Interrogatório

CNN – *Convolutional Neural Networks*

CRF – *Conditional Random Fields*

EI – Extração de Informações

EN – Entidade Nomeada

ER – Extração de Relações

IA – Inteligência Artificial

IDF – *Inverse Document Frequency*

IPL – Inquérito Policial

LSTM – *Long Short-Term Memory*

PLN – Processamento de Linguagem Natural

REN – Reconhecimento de Entidade Nomeada

RI – Recuperação de Informações

SRI – Sistemas de Recuperação de Informações

# SUMÁRIO

# 1. INTRODUÇÃO

As investigações policiais de combate ao crime de lavagem de dinheiro, previsto na Lei nº 9.613/98 e alterado pela Lei 12.683/12, ocupam um importante papel social com destaque na mídia, a exemplo da Operação Lava Jato, na medida em que enfrentam a corrupção e expõem o descaso com os recursos que deveriam ser revertidos para a sociedade, mas que estão sendo utilizados para o enriquecimento de poucos.

Nas palavras de [dLJ07], dentre os desafios encontrados na investigação do crime de lavagem de dinheiro, se destaca a dificuldade de visualizá-lo, pois não há uma vítima pontual e, tampouco, um único agente do delito. O concurso de pessoas é indispensável à consecução penal do delito, o que funciona como fator altamente complicador. Não se trata de uma simples co-autoria, tampouco de uma quadrilha ou bando, mas sim de uma complexa estrutura de pessoas organizadas em torno de um objetivo comum. Para tanto, tais agentes não precisam estar próximos ou agirem juntos.

A tecnologia tem sido amplamente utilizada pelas organizações criminosas para o cometimento destes crimes, especialmente em transações financeiras e ocultação de bens adquiridos ilicitamente.

Considerando a complexidade do delito em estudo e da tecnologia empregada, combinada ao grande volume de dados provenientes dos materiais produzidos e apreendidos em uma investigação policial, verifica-se a importância do estudo de técnicas de Processamento de Linguagem Natural (PLN) no auxílio às equipes de análise investigativa.

O PLN é um campo da Inteligência Artificial (IA) que objetiva tornar possível o entendimento da linguagem humana pelo computador, permitindo a obtenção de informações estruturadas de forma que elas possam ser utilizadas por máquinas para tarefas orientadas ao conhecimento.

Apesar da estrutura sintática e gramatical existentes nas linguagens naturais, o processamento e análise de textos são um grande desafio computacional, especialmente para a IA, isso porque essas linguagens são especialmente diferentes das linguagens de programação computacionais.

Este trabalho apresenta um estudo do Reconhecimento de Entidades Nomeadas (REN), que pode ser entendido como um processamento computacional cujo objetivo é identificar as entidades nomeadas e executar sua classificação, atribuindo uma categoria semântica para essas entidades.

No contexto das investigações policiais, as entidades nomeadas desempenham um papel fundamental na elucidação do crime, classificando possíveis autores, organizações envolvidas, e elementos que possam servir de prova do cometimento do delito.

## 1.1 Objetivo

O objetivo deste trabalho de dissertação é estudar técnicas de REN e PLN, identificando os recursos existentes para a língua portuguesa e as possibilidades de aplicações no domínio das investigações policiais de combate aos crimes de lavagem de dinheiro.

A atividade investigativa pode envolver a análise de diversos tipos de documentos, disponíveis nos mais variados formatos e estruturas, o que torna a tarefa de PLN ainda mais complexa. Sendo assim, para alcançar os objetivos propostos, foram selecionadas peças policiais fornecidas pela Polícia Federal do Rio Grande do Sul, produzidas em cartório durante a oitiva de testemunhas e indiciados, que apesar de serem textos livres, seguem um padrão mínimo de construção.

Dessa forma, a partir de um *corpus* selecionado, foram realizados experimentos práticos e analisados seus resultados, visando identificar a eficácia dos modelos de REN e técnicas de PLN para a tarefa proposta, podendo, a partir deste estudo, serem estendidos para tarefas mais complexas.

## 1.2 Organização do volume

Esta dissertação está dividida em sete capítulos: no capítulo dois, são apresentados os conceitos de investigação policial, a finalidade de um inquérito policial, as especificidades do crime de lavagem de dinheiro e as características das peças produzidas no cartório da polícia; no capítulo três, são apresentados os conceitos relacionados, necessários para a compreensão do trabalho; no capítulo quatro, apresentam-se os trabalhos relacionados; no capítulo cinco, uma pesquisa realizada com especialistas da Polícia Federal, objetivando identificar, do ponto de vista dos especialistas, as principais Entidades Nomeadas, relações entre entidades e verbos que possam auxiliar numa investigação policial de crimes de lavagem de dinheiro; no capítulo seis, são apresentados os experimentos realizados utilizando REN e outras técnicas de PLN; por fim, no capítulo sete, apresentam-se as considerações finais, com as limitações e trabalhos futuros.

# 2. INVESTIGAÇÕES POLICIAIS

A investigação policial é uma das atribuições de polícia judiciária, no Brasil realizada pelas polícias civis e pela Polícia Federal. O objetivo de uma investigação policial é identificar indícios que possam revelar a autoria e a materialidade de uma infração penal. Para tanto, os policiais utilizam técnicas presentes na doutrina policial, ensinadas nas academias de polícia. Contudo, não existe uma sequência de ações predeterminadas e rígidas a serem utilizadas na investigação policial, dada a dinâmica envolvida.

Segundo [FJD07], dado o complexo cenário das organizações criminosas, é fundamental que as técnicas de investigação policial também passem a ser complexas e sofisticadas, como pré-requisito básico para sua necessária efetividade. O excesso de informação, ao contrário do que possa parecer, desafia as limitações normais do intelecto humano, exigindo mais que uma percepção intuitiva e um “trabalho artesanal”.

Assim, além das exigências específicas para o exercício profissional em uma organização policial, verifica-se a necessidade da utilização de tecnologias avançadas que incrementem, em um curto espaço de tempo, a capacidade cognitiva dos policiais.

## 2.1 Inquérito Policial

O Inquérito Policial (IPL) é um procedimento investigatório, presidido por um delegado de polícia, instaurado em razão da prática de uma infração penal, composto por uma série de diligências, que tem como objetivo obter elementos de prova para que o titular da ação possa propô-la contra o criminoso [RG18]. Em suma, quando é cometido um delito, deve o Estado, por intermédio da polícia judiciária (Polícia Civil ou Federal), buscar provas iniciais acerca da autoria e da materialidade, para apresentá-las ao titular da ação penal (Ministério Público ou ofendido).

Para iniciar uma ação penal pública é necessário que o membro do Ministério Público ofereça uma denúncia, peça inaugural que contém sua convicção diante dos elementos contidos no inquérito policial, ou mediante outras peças informativas, que indiquem a existência de fato criminoso e indícios de autoria.

Pesquisa desenvolvida pela Escola Superior do Ministério Público da União (ESMPU) [MZR16] aponta uma baixa taxa de denúncias para os crimes financeiros, na ordem de 33% para os inquéritos da Polícia Federal no Distrito Federal, 46,15% no estado de Pernambuco, 14,7% para o estado do Paraná e 10,70% para o estado de São Paulo. A persecução penal dos crimes de lavagem de ativos indica taxas inferiores a 10% em São Paulo e no Paraná, enquanto no Distrito Federal chega a 16% e 50% em Pernambuco. Ainda segundo a pes-

quisa, em uma análise quantitativa, identificou-se um número elevado de arquivamentos em relação aos inquéritos concluídos pela Polícia Federal.

Em que pese a relevância dos números expostos, vale ressaltar que o índice de oferecimento de denúncias, por si só, não reflete a efetividade das investigações e inquéritos policiais, uma vez que a investigação pode ter sido efetiva porém concluído pela inexistência de crime, dando causa ao não oferecimento da denúncia.

Segundo ensinamentos de [FS07], a finalidade do inquérito policial é a colheita de elementos que auxiliam a elucidação da autoria e da materialidade de determinada infração penal, visando subsidiar futura e eventual ação penal pública, quando proposta pelo Ministério Público, ou privada, nos crimes de ação penal privada.

O inquérito policial é um conjunto de peças composto de laudos técnicos, depoimentos tomados em cartório, informações e de um relatório, juridicamente orientado, que aponta as conclusões quanto à autoria e materialidade delitiva [Mis11].

Apesar da evolução dos meios digitais, até o presente momento, o inquérito policial possui forma escrita, não lhe conferindo a forma oral, como se verifica nos ensinamentos de [Tou18]. As peças elaboradas no procedimento policial têm por finalidade subsidiar a ação penal e devem extirpar a interpretação da caligrafia dos escrivães, evitando riscos de erros e borrões, que podem levar o leitor a uma interpretação equivocada.

O fato do IPL ser escrito reafirma a justificativa do estudo do Processamento de Linguagem Natural em textos.

## 2.2 Crime de lavagem de dinheiro

Também conhecido como branqueamento de capitais, o crime de lavagem de dinheiro está previsto na Lei nº 9.613/98, alterada pela Lei nº 12.683/12, em resposta ao apelo internacional da Convenção contra o Tráfico Ilícito de Entorpecentes e Substâncias Psicotrópicas, conhecida como Convenção de Viena de 1988. Inicialmente a Lei nº 9.613/98 associou a lavagem de dinheiro aos crimes de tráfico de drogas, numa segunda fase, a lei ampliou o conceito de crime antecedente, envolvendo um rol maior de crimes. Atualmente, com o advento da Lei nº 12.683/12, o crime antecedente foi ampliado para qualquer infração penal, assim entendida toda ação que visa “ocultar ou dissimular a natureza, origem, localização, disposição, movimentação ou propriedade de bens, direitos ou valores provenientes, direta ou indiretamente, de infração penal”.

Segundo [Bad16], lavagem de dinheiro é o ato ou a sequência de atos praticados para mascarar a natureza, origem, localização, disposição, movimentação ou propriedade de bens, valores e direitos de origem delitiva ou contravencional, com o escopo último de reinseri-los na economia formal com aparência de licitude.

O crime de lavagem de dinheiro está sistematizado em três fases [DE 11]:

- **colocação**: alocação ou movimentação de ativos, ou ainda, mudança na forma de apresentação destes, dificultando ou impossibilitando seu rastreamento.
- **ocultação**: simulação de uma origem lícita para os recursos que têm uma fonte espúria, conferindo-lhes uma nova fonte.
- **integração**: distanciamento do recurso criminoso da origem sob ponto de vista pessoal, vinculando-o a pessoas (físicas ou jurídicas) “limpas”, sem relação ou ligação com os delitos antecedentes.

Para [Cas17], o crime de lavagem de dinheiro é complexo, possui caráter transnacional, envolve uma série de etapas e seus agentes estão sempre aprimorando o *modus operandi*, de forma a dificultar a persecução criminal. Esse ilícito está sempre acompanhado de outros tipos penais e a grande quantidade de recursos envolvidos geram maior preocupação hoje nas nações e despertaram na comunidade internacional a necessidade de combatê-lo.

## 2.3 Peças Policiais

A fim de subsidiar a convicção do Ministério Público sobre a existência de elementos suficientes para dar início à ação penal, a polícia judiciária promove a oitivas de suspeitos e testemunhas, reduzindo a termo seus depoimentos. Dentre as peças produzidas estão o Termo de Depoimento, Termo de Declaração e Auto de Qualificação e Interrogatório.

- **Auto de Qualificação e Interrogatório** (Figura 2.2): peça policial lavrada para a oitiva do indivíduo que é indiciado, quando há indícios suficientes que apontam ser este o autor do crime investigado. O interrogado não tem o dever de dizer a verdade, sob pena de produzir prova contra ele mesmo.
- **Termo de Declaração** (Figura 2.1): peça policial utilizada para a oitiva de alguém que se presume ser o autor do crime investigado, mas ainda há dúvidas quanto à autoria. O declarante não tem a obrigatoriedade de dizer a verdade, sob pena de produzir prova contra ele mesmo.
- **Termo de Depoimento** (Figura 2.1): peça policial lavrada para a oitiva de testemunhas. A estrutura é a mesma do Termo de Declaração. Das três peças citadas, esta é a única peça em que a pessoa se compromete a dizer a verdade, sob pena de incorrer em crime de falso testemunho.

As informações produzidas por meio de depoimentos desempenham um papel fundamental na elucidação criminal, uma vez que possibilitam a demonstração da verdade por percepções sensoriais como a visão, audição, e sistemas sinestésicos. [Mir03] explica a importância da prova testemunhal para o processo penal:

> Como a prova, no processo, tem por fim demonstrar a verdade de determinados fatos, é muitas vezes indispensável que sejam ouvidas as pessoas que os presenciaram, no todo ou ao menos em parte. Essas pessoas passam a ser testemunhas do fato. No sentido legal, testemunha é a pessoa que, perante o juiz, declara o que sabe acerca dos fatos sobre os quais se litiga no processo penal ou as que são chamadas a depor, perante o juiz, sobre suas percepções sensoriais a respeito dos fatos imputados ao acusado. Isto porque, o conhecimento da testemunha a respeito dos acontecimentos lhe é fornecido pelos seus sentidos, em especial a visão e a audição, não se podendo excluir, também, em determinadas hipóteses, o paladar, o olfato e o tato. (Mirabete, 2003, p. 292)

Para a realização dos experimentos presentes neste trabalho, foram selecionadas 15 peças policiais produzidas pela Superintendência da Polícia Federal no Rio Grande do Sul, entre os anos de 2010 e 2011, todas com investigações encerradas. A opção por textos mais antigos se deu em razão do sigilo das informações mais recentes. Foi firmado um Termo de Compromisso de Confidencialidade com a Polícia Federal que permite a utilização das informações para publicações científicas desde que não constem nomes e qualificações das pessoas e empresas envolvidas.

As peças selecionadas possuem uma média de 928 palavras. Estão divididas em 05 Autos de Qualificação e Interrogatório, 05 Termos de Depoimento e 05 Termos de Declaração.

TERMO DE DECLARAÇÕES DE
[NOME DO DECLARANTE]:

Ao$_{(s)}$ 13 dia(s) do mês de junho de 2011, nesta Superintendência Regional de Polícia Federal, em Porto Alegre/RS, onde se encontrava [NOME DO DELEGADO], Delegado de Polícia Federal, compareceu [NOME DO DECLARANTE], sexo masculino, nacionalidade [NACIONALIDADE], casado(a), filho(a) de [NOME DO PAI] e [NOME DA MÃE], nascido(a) aos [DATA DE NASCIMENTO], natural de [CIDADE], instrução ensino médio ou técnico profissional, profissão Desempregado(a), documento de identidade n° [Nº DA IDENTIDADE], CPF [Nº DO CPF], residente na(o) [ENDEREÇO], fone [TELEFONE], celular [CELULAR]. Inquirido a respeito dos fatos, RESPONDEU: **QUE,** primeiramente, gostaria de registrar que, inobstante todo o interesse que tenha de colaborar com a investigação, os fatos ora em questão já se passaram há muitos anos e, assim, sua memória já não guarda mais tantos detalhes; **QUE** tem a dizer quanto aos fatos é que muitos dos negócios que, na época dos fatos, envolveram seu nome, foram feitos na realidade em auxílio a outros operadores

Figura 2.1 – Exemplo de Termo de Declarações (os dados foram ocultados)

A Figura 2.1 apresenta um exemplo de Termo de Declarações, idêntico ao Termo de Depoimento que, como dito anteriormente, difere do primeiro por ser a transcrição da

oitiva de uma testemunha que possui o dever de dizer a verdade, sob pena de incorrer no crime de falso testemunho.

Os dados pessoais foram anonimizados em razão do Termo de Confidencialidade firmado com a Polícia Federal, visando a proteção das informações pessoais e das empresas constantes nos documentos.

AUTO DE QUALIFICAÇÃO E INTERROGATÓRIO
DE: **[NOME DO INTERROGADO]**

Ao(s) 27 dia(s) do mês de setembro de 2011, nesta Superintendência Regional do Departamento de Polícia Federal, em Porto Alegre/RS, onde se encontrava [NOME DO DELEGADO], Delegado de Polícia Federal, pelo(a) mesmo(a) foi determinado que se formalizasse a qualificação do(a) indiciado(a), o(a) qual RESPONDEU:
NOME: **[NOME DO INTERROGADO]**
ALCUNHA: não possui
NACIONALIDADE: brasileira
ESTADO CIVIL: casado(a)
PAI: [NOME DO PAI]
MÃE: [NOME DA MÃE]
DATA DE NASCIMENTO: [DATA]
NATURALIDADE: [CIDADE]
PROFISSÃO: [PROFISSÃO]
INSTRUÇÃO: Terceiro Grau Completo
DOCUMENTO DE IDENTIDADE: [RG]
CPF: [CPF]
RESIDÊNCIA: [ENDEREÇO]
ENDEREÇO COMERCIAL: [ENDEREÇO]
Cientificado(a) das imputações que lhe são feitas e de seus direitos constitucionais, inclusive o de permanecer calado(a), PERGUNTADO Qual a profissão e/ou atividade profissional desempenhada pelo interrogado? Qual a remuneração mensal média que recebe nessas atividades? RESPONDEU **QUE** irá exercer o seu direito de permanecer calado; PERGUNTADO Onde reside? Desde quando? Reside em imóvel próprio, alugado, cedido? RESPONDEU **QUE** irá exercer o seu direito de permanecer calado; PERGUNTADO se conhece os imóveis localizados nas ruas [ENDEREÇO]. É ou foi proprietário destes imóveis? Em caso positivo, quando e com que recursos os adquiriu? Qual a finalidade da aquisição destes imóveis? RESPONDEU **QUE** irá exercer o seu

Figura 2.2 – Exemplo de Auto de Qualificação e Interrogatório (os dados foram ocultados)

A primeira parte do texto, que antecede o primeiro “QUE”, é a parte de qualificação e segue um padrão utilizado em todas as peças produzidas nas oitivas policiais. As sentenças são iniciadas pela palavra “QUE” (com letras maiúsculas e negritada) e finalizadas por um ponto-e-vírgula. A transcrição é feita a partir das respostas dadas pelo interrogado, declarante ou depoente, às perguntas formuladas pela autoridade policial.

Os nomes de pessoas e empresas (EN pessoas e organização) são comumente escritas em caixa alta. O fato de serem textos livres podem gerar distorções neste padrão.

Ressalta-se que estes documentos representam uma pequena parte do que é analisado por uma equipe investigativa, uma vez que, dada a complexidade de uma investigação, o conjunto probatório possui outras diversas fontes, como é o caso de laudos periciais, materiais apreendidos e outros arrecadados durante a fase de deflagração de uma operação policial.

# 3. CONCEITOS RELACIONADOS

Segundo [VL10], Processamento de Linguagem Natural (PLN) é uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais.

O que torna o PLN uma tarefa relevante e complexo para a computação é o fato da linguagem natural, diferentemente da linguagem computacional, ser rica em ambiguidades.

Os primeiros trabalhos realizados na área de PLN possuem registros na década de 1940. Nesta época os desafios estavam voltados para a tradução automática de documentos relacionados à guerra, como consta no trabalho de [Wea55]. Nas décadas seguintes, os trabalhos de PLN passaram do entendimento do discurso para as questões semânticas.

O crescimento da internet aumentou significativamente o volume de informações, impulsionando a utilização de métodos estatísticos e linguísticos para o reconhecimento e geração automática da linguagem natural.

## 3.1 Extração de Informações

A Extração de Informações (EI) ou Recuperação de Informações (RI) é um campo da Ciência da Computação responsável pela identificação de informações relevantes em textos livres contidos em documentos. Para [Sou06], um dos desafios do processo de recuperação de informações é a predição de quais documentos e informações são relevantes e quais devem ser descartados, baseados em heurísticas ou critérios previamente definidos.

Uma das ferramentas mais importantes para auxiliar os Sistemas de Recuperação de Informações (SRI) é o índice, que consiste em uma coleção de termos que indicam o local onde a informação pode ser localizada. Na Figura 3.1, [Car04] apresenta uma forma sintetizada do funcionamento de um SRI.

Verifica-se que o processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário. Os índices construídos para uma coleção de documentos são usados para acelerar esta tarefa.

## 3.2 Reconhecimento de Entidades Nomeadas

Entidades Nomeadas (EN) compreendem termos que visam restringir designadores rígidos, que incluem nomes próprios e certos termos naturais, como espécies e subs-

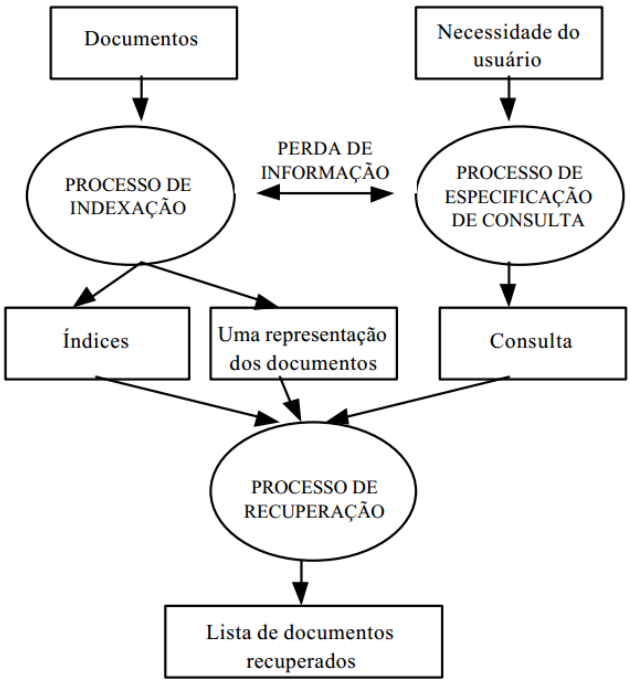


Figura 3.1 – Componentes de um Sistema de Recuperação de Informações.

tâncias biológicas, além de expressões temporais e algumas expressões numéricas, como quantias em dinheiro e outros tipos de unidades [NS07].

Reconhecimento de Entidades Nomeadas (REN), segundo [Moh14], é o problema de localizar e categorizar nomes importantes e nomes próprios em um texto livre. Por exemplo, em notícias, nomes de pessoas, organizações e locais são normalmente importantes.

Originalmente, o REN se limitava à extração de nomes próprios relacionados às notícias, como nomes de pessoas, organizações e locais. Com a expansão do PLN em outros domínios, essas poucas classes de entidades nomeadas tradicionais se tornaram insuficientes. Por exemplo, para um artigo sobre ciência ou tecnologia, as três classes tradicionais não são suficientes e outras classes de entidades nomeadas precisam ser consideradas. Além disso, as entidades nomeadas não devem se limitar aos nomes próprios. Em certas áreas de estudos como a física nuclear, pode-se destacar termos como prótons ou urânio como entidades nomeadas. Assim, apesar do foco comum nas classes pessoa, localização e organização, pode-se dizer que o REN abrange a extração de todas as entidades importantes em um determinado contexto.

As primeiras abordagens do Reconhecimento de Entidades Nomeadas eram baseadas principalmente em regras, os sistemas eram relativamente precisos, geralmente com baixa cobertura, mas funcionam bem em domínios estreitos. Seu desempenho geralmente depende de quão abrangentes são as regras e os léxicos.

Posteriormente, com a crescente popularidade dos métodos estatísticos de PLN, juntamente com a expansão dos recursos de dados disponíveis, as pesquisas de REN foram

direcionadas para o uso de métodos estatísticos, reduzindo o esforço humano necessário para a construção de conjuntos de regras.

## 3.3 *Corpus*

Segundo a definição de [PML96], *corpus* é uma coletânea de porções de linguagem que são selecionadas e organizadas de acordo com critérios linguísticos explícitos, a fim de serem usadas como uma amostra da linguagem.

No ano de 1999 comemorou-se o aniversário de 35 anos da criação do primeiro *corpus* linguístico eletrônico, o Corpus Brown [KF67]. Lançado em 1964, o Brown University Standard Corpus of Present-Day American English, continha uma quantidade invejável de dados para a época: um milhão de palavras.

Embora os computadores armazenem inúmeras informações relevantes para os usuários, eles não estão prontamente aptos a entender a linguagem em si contida nos arquivos. Assim, iniciaram-se pesquisas sobre corpora anotados sintaticamente, ou seja, conjuntos de textos sobre um domínio de conhecimento onde cada uma das suas palavras foram identificadas segundo sua função sintática.

Segundo [PS12], a linguística teórica e computacional está focada em desvendar a natureza mais profunda da linguagem e capturar as propriedades computacionais das estruturas linguísticas. As tecnologias de linguagem humana tentam adotar esses *insights* e transformá-los em programas funcionais de alto desempenho que podem afetar a forma como interagimos com computadores usando a linguagem.

Em razão do aumento de pessoas usando a internet, a quantidade de dados linguísticos disponíveis para pesquisadores aumentou significativamente, permitindo que os problemas de modelagem linguística sejam vistos como tarefas de aprendizado de máquina, em vez de limitados às quantidades relativamente pequenas de dados que os humanos são capazes de processar por conta própria. No entanto, não é suficiente fornecer a um computador uma grande quantidade de dados e esperar que ele aprenda a falar - os dados precisam ser preparados de tal maneira que o computador possa encontrar mais facilmente padrões e inferências. Isso geralmente é feito adicionando metadados relevantes a um conjunto de dados.

Qualquer *tag* de metadados usada para marcar elementos do conjunto de dados é chamada de anotação sobre a entrada. No entanto, para que os algoritmos aprendam com eficiência e eficácia, a anotação feita nos dados deve ser precisa e relevante para a tarefa que a máquina está sendo solicitada a executar. Por essa razão, a disciplina de anotação de linguagem é um elo crítico no desenvolvimento de tecnologias inteligentes de linguagem humana.

Conjuntos de dados de linguagem natural são referidos como corpora, e um único conjunto de dados anotados é chamado de um *corpus* anotado. O *corpus* anotado pode ser usado para treinar algoritmos de aprendizado de máquina. Dar a um algoritmo muita informação pode atrasá-lo e levar a resultados imprecisos. É importante pensar com cuidado sobre o que se está tentando realizar e quais informações são mais relevantes para esta tarefa.

## 3.4 Pré-processamento do texto

A etapa que antecede o pré-processamento do texto é a coleta de dados ou documentos, responsável pela seleção e recuperação de documentos relevantes ao domínio do conhecimento a ser extraído.

Considerando que a primeira etapa do processo tenha sido cumprida, ou seja, os documentos estejam disponíveis, é necessário realizar o pré-processamento. Segundo [Ima01], o pré-processamento de textos consiste em um conjunto de ações realizadas sobre alguma coleção de textos com o objetivo de fazer com que estes passem a ser estruturados em uma representação atributo-valor que possa ser manipulada pelos métodos de extração de conhecimento.

A fase de preparação dos textos possui grande relevância para a tarefa de PLN, grande parte dos resultados obtidos ao fim do processo pode depender de quão bem a representação atributo-valor descreve o contexto. Como nem todas as técnicas são adequadas a todo tipo de domínio, podem ser necessários vários experimentos a fim de obter uma representação satisfatória, o que pode exigir grande parte do tempo de todo o processo.

### 3.4.1 Tokenização

A tokenização é a primeira fase do pré-processamento de textos, sendo ela uma etapa crucial na segmentação da informação. Utiliza-se a técnica de separação de palavras a partir de um caracter pré-definido, geralmente o espaço.

Utilizando-se o espaço como delimitador, a frase “Foi então advertido da obrigatoriedade de comunicação de eventuais mudanças de endereço em face das prescrições do Art. 224 do CPP” seria dividida em 21 *tokens*: “Foi” - “então” - “advertido” - “da” - “obrigatoriedade” - “de” - “comunicação” - “de” - “eventuais” - “mudanças” - “de” - “endereço” - “em” - “face” - “das” - “prescrições” - “do” - “Art.” - “224” - “do” - “CPP”.

Este ainda é um estágio básico de processamento, com a possibilidade do *token* segmentado possuir a mesma interpretação ou relevância em contextos diferentes, causando o problema da ambiguidade.

### 3.4.2 *Stopwords*

Uma lista de *stopwords* ou dicionário negativo é uma técnica usada no processamento de textos para filtrar palavras que resultariam em termos de índices ruins. Tradicionalmente, as listas de paradas devem incluir apenas as palavras que ocorrem com mais frequência e que não possuem relevância para o trabalho de PLN.

É comum o uso de um conjunto de *stopwords* que agrega os termos da língua, a exemplo de artigos, pronomes, preposições e conjunções. É importante observar, ainda, que o as *stopwords* são sensíveis ao contexto, por exemplo, ao retirar o termo “A” num contexto de saúde, pode-se excluir indevidamente referências à “vitamina A”.

O desafio desta tarefa consiste na elaboração de listas com *stopwords* eficientes na filtragem das palavras que ocorrem com maior frequência e semanticamente neutras na literatura geral, como é demonstrado no trabalho realizado por [Fox89].

### 3.4.3 *Bag-of-words*

Visando estruturar os textos para sua análise e manipulação, uma das tarefas relevantes é a escolha da estrutura para representação dos documentos. Uma das possíveis abordagens para a estruturação de documentos é denominada *bag-of-words*.

Como explica [Say07], esta tarefa se utiliza de matrizes termo-documento, onde cada linha representa um documento e cada coluna representa um termo presente em cada documento. O valor da célula representa o peso que determinado termo possui no documento analisado, conforme Tabela 3.1.

Tabela 3.1 – Estrutura genérica de uma matriz termo-documento

| | $termo_1$ | $termo_2$ | $termo_3$ | $termo_4$ | ... | $termo_t$ |
|---|---|---|---|---|---|---|
| $doc_1$ | $peso_{11}$ | $peso_{12}$ | $peso_{13}$ | $peso_{14}$ | ... | $peso_{1t}$ |
| $doc_2$ | $peso_{21}$ | $peso_{22}$ | $peso_{23}$ | $peso_{24}$ | ... | $peso_{2t}$ |
| ... | ... | ... | ... | ... | ... | ... |
| $doc_n$ | $peso_{n1}$ | $peso_{n2}$ | $peso_{n3}$ | $peso_{n4}$ | ... | $peso_{nt}$ |

Existem diversas abordagens para a definição do peso e consequente relevância de um termo para o documento, recebendo valor zero quando o termo não estiver presente.

Um dos desafios que o uso de *bag-of-words* apresenta é a grande dimensão das matrizes, podendo ser reduzida com o uso eficaz de outras técnicas, como as *stopwords*, mencionadas na Seção 3.4.2.

## 3.5 Visualização de informações

A visualização de informações é uma área de aplicação de técnicas de computação gráfica, geralmente interativas, que visa auxiliar o processo de análise e compreensão de um conjunto de dados, através de representações gráficas manipuláveis. Uma técnica de visualização é baseada numa representação visual e em mecanismos de interação que possibilitam ao usuário manipular essa representação de modo a melhor compreender o conjunto de dados ali representados [FCLC01].

A visualização é muito mais que uma simples amostragem de dados, é uma forma de facilitar a leitura dos mesmos, possibilitando a análise por meio de cruzamentos das variáveis disponíveis. A partir da análise das informações, através da percepção visual e dos gráficos, é possível extrair e gerar conhecimento.

Com a ajuda dos sentidos, nesse caso a visão, a cognição, que é o processo humano de aquisição e uso do conhecimento, se torna mais fácil, constatando padrões e características visuais presentes nas imagens, o que seria difícil perceber observando simplesmente os dados em sua forma bruta [Nov10].

Uma das técnicas de visualização muito utilizada em dados textuais é a nuvem de palavras, proporcionando uma forma simples e útil para uma análise rápida do contexto.

### 3.5.1 Nuvem de Palavras

As nuvens de palavras surgiram como um método de visualização simples e visualmente atraente para textos. Elas são utilizadas em vários contextos como um meio de fornecer uma visão geral, ressaltando as palavras que aparecem com maior frequência. Conforme se depreende do trabalho de [HLLE14], o método de visualização por nuvem de palavras, aliado às técnicas avançadas de processamento de linguagem natural, pode ser utilizado com eficácia para resolver tarefas de análise de textos, servindo como um ponto de partida para uma análise mais profunda.

Uma desvantagem aparente das nuvens de palavras é que a diferença na frequência entre os termos, julgada de acordo com o tamanho da fonte, pode dar uma impressão falsa sobre a verdadeira taxa de contagem de frequência dos termos. Mostrar os valores absolutos de frequência para os usuários permite que eles corrijam facilmente impressões

falsas. Em contrapartida, proporciona uma visão macro dos dados facilitando a tomada de decisão por parte do analista.

### 3.5.2 Escala Multidimensional (MDS)

Originado em trabalhos da área de psicofísica, a Escala Multidimensional (MDS) pode ser definida como um mapeamento injetivo entre objetos pertencentes a um espaço m-dimensional em pontos em um outro espaço p-dimensional, buscando-se preservar relações de distância [Pau08].

A técnica MDS é usada para representar espacialmente, em 2D ou 3D, uma matriz de proximidades (semelhança ou dissemelhança) entre uma série de objetos de modo que possam ser mais facilmente visualizados.

O objetivo da MDS é encontrar uma configuração de pontos de tal forma que a distância entre os mesmos sejam relacionados às similaridades entre os objetos por alguma função de transformação.

# 4. TRABALHOS RELACIONADOS

Com o objetivo de auxiliar investigadores criminais na análise de grandes quantidades de informação textual de maneira mais eficiente e rápida, em [vBKLK16], os autores apresentam uma ferramenta denominada LES, desenvolvida com técnicas de PLN. Fora avaliado o desempenho da ferramenta com diferentes métricas e apresentados os resultados experimentais com conjuntos de dados grandes e complexos. Os autores identificaram algumas dificuldades enfrentadas pelos analistas no uso de sistemas forenses para a investigação, como: tempo gasto para processar todos os dados apreendidos; os softwares forenses são incapazes de lidar com a enorme quantidade de dados de uma investigação; falha de software ao consultar os bancos de dados; tempo de espera inaceitável ao realizar uma consulta; muitos *hits* de busca para fazer a análise da evidência humanamente possível em muitos casos; e muita abordagem técnica na interface. Os principais ganhos no uso do sistema foram apontados como sendo: melhoria no tempo de processamento de dados para manipular grandes quantidades de dados, melhoria no tempo de análise de conjuntos de dados complexos, e permitir que os usuários finais executem tarefas complexas com uma interface muito simples.

Técnicas de aprendizado de máquinas têm sido muito utilizadas em PLN, como exposto no artigo [dAV14] em que as autoras utilizam o algoritmo *Conditional Random Fields* (CRF) para a tarefa de REN em corpora da língua portuguesa e avaliam comparativamente o desempenho desse método com outros sistemas, tendo como base o corpus do HAREM (avaliação conjunta na área do Reconhecimento de Entidades Mencionadas).

Uma técnica relevante no PLN, que complementa a tarefa de REN, é a Extração de Relações entre entidades nomeadas (ER). O artigo [CMV16] aborda a extração e estruturação de relações abertas entre entidades nomeadas. Foi aplicado o modelo CRF para a extração de qualquer descritor de relações expressando qualquer tipo de relação entre um par de entidades nomeadas (categorias pessoa, lugar e organização).

No trabalho desenvolvido por [Pir15], são abordadas as formas aberta e convencional de Extração de Informação em textos livres, onde os autores fazem uma comparação dos resultados das ferramentas, indicando uma maior eficiência nas ferramentas abertas. Por outro lado, os resultados apresentaram melhor precisão na técnica convencional, que depende de anotação. Dessa forma, os autores indicam uma análise em relação aos objetivos pretendidos para a utilização da técnica mais apropriada.

[KIL08] propõe o desenvolvimento de uma ferramenta para coletar os depoimentos de vítimas e testemunhas de crimes, utilizando PLN para a extração de informações relevantes, visando auxiliar a análise investigativa. Os autores acreditam que as vítimas e testemunhas se sentem inibidas ao relatar os fatos à uma autoridade policial, e que se os

dados fossem relatados em texto corrido, de forma anônima, poderia conter maior detalhe e canalizar para uma maior elucidação dos crimes por parte das equipes investigativas.

Seis anos após o trabalho anterior, os autores publicaram um novo trabalho. Em [KL13] foi identificada a dificuldade em lidar com os depoimentos anônimos de vítimas e testemunhas de crimes. Para solucionar o problema na identificação de correlação entre os relatos anônimos e determinados crimes, os autores desenvolveram um sistema de suporte à decisões, combinando técnicas de PLN, medidas de similaridade e aprendizado de máquina, com o uso de um classificador Naive Bayes, para apoiar as análises, classificar e identificar a relação entre os relatórios criminais.

Em [JC11] foi proposta uma análise à aplicação do REN no campo da computação forense, em especial nas tarefas associadas ao exame de mídias apreendidas pela Polícia Federal da Bahia, demonstrando a contribuição para a redução do tempo investido na etapa de análise de conteúdo das mídias apreendidas e para a revelação de informações latentes de nomes de pessoas e organizações contidas nessas mídias. Os resultados do trabalho demonstraram que a utilização do REN auxilia na seleção de arquivos relevantes, reduzindo a quantidade de arquivos pendentes de análise manual por parte da equipe de investigação.

Com o objetivo de explorar e contribuir para o estado da arte das técnicas de Reconhecimento de Entidades Nomeadas (REN) e Extração de Relações (RE) em português, o Fórum de Avaliação de Idiomas Ibéricos (IberLEF) propôs três tarefas independentes. A primeira tarefa, de REN, consistiu em identificar nomes próprios em um determinado texto e classificá-los em uma das muitas categorias relevantes ou dentro de uma categoria padrão conhecida como Diversos. A segunda, de RE, envolveu a extração automática de qualquer descritor de relação que expressasse qualquer tipo de relação entre um par de categorias de Entidades Nomeadas das classes Pessoa, Local e Organização nos textos em português. A terceira tarefa consistiu numa generalização da segunda, removendo o requisito das entidades nomeadas no texto, o que significa que considerou qualquer relação entre duas frases substantivas (*Noun Phrases*-NP). O *dataset* jurídico aplicado nas tarefas foi o mesmo aplicado neste trabalho e foi uma contribuição deste autor. O trabalho de [CNC+19] descreve os resultados das seis ferramentas participantes e constatou que os resultados do conjunto de dados policial, em particular, mostraram uma notável diferença entre as abordagens baseadas nas Redes Neurais e as que não eram. Outros trabalhos, como observado por [PAS+19] e [dCdSS19], relatam o melhor desempenho das ferramentas para o *corpus* policial quando comparado ao clínico, podendo ser atribuído à padronização e ao vocabulário utilizado nas peças policiais.

# 5. PESQUISA COM ESPECIALISTAS DA POLÍCIA FEDERAL

A fim de identificar quais são as principais informações presentes nas peças policiais que poderiam auxiliar a análise e a identificação de autoria e materialidade dos crimes investigados, foi realizada uma pesquisa, por meio de formulário digital, com especialistas da Polícia Federal que atuam diretamente em investigações de crimes financeiros.

A pesquisa foi respondida por 24 policiais, sendo 12 Delegados de Polícia Federal, 03 Peritos Criminais Federais, 06 Agentes de Polícia Federal e 03 Escrivães de Polícia Federal, todos atuantes nos setores de Combate a Crimes Financeiros.

Em conversas preliminares com os participantes, foram identificadas as principais Entidades Nomeadas, as relações entre elas e os verbos mais utilizados nas investigações. Foi elaborado, então, um formulário digital solicitando aos policiais que ordenassem as respostas conforme a relevância.

As principais classes de Entidades Nomeadas identificadas na pesquisa estão apresentadas na Tabela 5.1, ordenadas por relevância.

Tabela 5.1 – Pesquisa: principais classes de Entidades Nomeadas

| Ordem | Entidade Nomeada |
|---|---|
| 1 | Pessoas físicas |
| 2 | Contas bancárias, títulos e valores mobiliários |
| 3 | Lugares e endereços |
| 4 | Documentos e contratos públicos |
| 5 | Documentos e contratos privados |
| 6 | Organizações privadas |
| 7 | Cargo ou função |
| 8 | Organizações públicas |
| 9 | Bens imóveis |
| 10 | Período (tempo) |
| 11 | Veículos |
| 12 | Moeda nacional |
| 13 | Moeada estrangeira |
| 14 | Jóias, pedras e metais preciosos |
| 15 | Obras de arte |

Percebe-se que a entidade com maior relevância para as investigações de crimes financeiros, segundo a pesquisa, é “pessoa física”. As pessoas jurídicas “organizações privadas” e “organizações públicas” ficaram na 6ª e 8ª posição, respectivamente. A constatação é coerente com o objetivo penal, que é a identificação da autoria delitiva, uma vez que

pessoa jurídica não pode cometer crime financeiro, como exposto por [Gom95] que defende a punição à pessoa jurídica quando beneficiada pelo crime, porém em caráter sancionador, não penal.

As principais relações entre as entidades identificadas na pesquisa foram as relacionadas na Tabela 5.2.

Tabela 5.2 – Pesquisa: principais relações entre Entidades Nomeadas

| Ordem | Relação |
|---|---|
| 1 | É sócio somente no papel (laranja, p. ex.) |
| 2 | É proprietário |
| 3 | É sócio legal (em um contrato social, p. ex.) |
| 4 | É procurador |
| 5 | Outorgou procuração |
| 6 | É servidor, funcionário ou empregado |
| 7 | É ou foi indicado (para um cargo ou função) |
| 8 | É cônjuge ou companheiro(a) |
| 9 | É parceiro ou auxiliar (conscientemente) |
| 10 | É usuário |
| 11 | Conhece |
| 12 | É auxiliar (inconscientemente) ou foi usado |
| 13 | É amigo |
| 14 | É credor |
| 15 | É devedor |
| 16 | É inimigo |
| 17 | É vendedor |
| 18 | Desconhece |
| 19 | Já ouviu falar |

Esta segunda parte da pesquisa revela a importância de se analisar a relação entre as entidades. As principais relações identificadas dizem respeito a entidades distintas (pessoas físicas e pessoas jurídicas). Apesar das pessoas físicas possuírem maior relevância para a persecução penal, é por meio das pessoas jurídicas que a maioria dos crimes financeiros são consumados, justificando a importância desta relação.

E na Tabela 5.3 são apresentados os verbos que, segundo a pesquisa, possuem maior relevância para as investigações de crimes financeiros.

A pesquisa serviu como ponto de partida, apontando para a viabilidade de se trabalhar inicialmente com as classes de entidades pessoa, localização e organização, sendo estas classes apontadas como relevantes para o domínio estudado e presentes na maioria dos modelos computacionais que trabalham com REN.

Ainda em relação às Entidades Nomeadas, a pesquisa revelou a importância de algumas classes consideradas relevantes para um trabalho futuro, como é o caso das “contas bancárias, títulos e valores mobiliários”.

Quanto aos verbos e aos termos que indicam relação entre as Entidades Nomeadas, ficam como sugestão para trabalhos futuros que pretendam explorar a Relação das Entidades (RE), tarefa que, em soluções fechadas, possui alta dependência do idioma e domínio, conforme [SC14].

Tabela 5.3 – Pesquisa: principais verbos

| Ordem | Verbo |
|---|---|
| 1 | Ocultar |
| 2 | Subornar |
| 3 | Mandar, determinar ou influenciar |
| 4 | Auferir ganho, lucrar ou aproveitar |
| 5 | Adquirir |
| 6 | Transferir ou remeter |
| 7 | Pedir, solicitar, exigir ou receber |
| 8 | Comprar |
| 9 | Corromper ou enganar |
| 10 | Assinar |
| 11 | Usar (sem ser proprietário) |
| 12 | Adulterar (uma coisa) |
| 13 | Vender |
| 14 | Prejudicar, eliminar ou destruir |
| 15 | Subtrair |
| 16 | Doar |
| 17 | Intermediar |
| 18 | Emprestar ou ceder gratuitamente |
| 19 | Alugar |

# 6. EXPERIMENTOS

Os experimentos realizados neste trabalho visam identificar e analisar possíveis soluções utilizadas no Processamento de Linguagem Natural (PLN), objetivando automatizar e auxiliar o trabalho de análise nas investigações policiais.

A partir de um *corpus* policial, foram realizados dois experimentos na tarefa de Reconhecimento de Entidades Nomeadas (REN), além da aplicação de outras técnicas de PLN como forma de expandir as possibilidades aplicáveis.

## 6.1 Recursos

Esta seção apresenta os recursos (*corpus* e modelos) necessários para a realização dos experimentos.

### 6.1.1 *Corpus*

Para a realização dos experimentos, foram selecionadas 30 peças policiais, sendo 10 Autos de Qualificação e Interrogatório, 10 Termos de Declaração e 10 Termos de Depoimento. As peças selecionadas foram produzidas pela Superintendência da Polícia Federal no Rio Grande do Sul, entre os anos de 2010 e 2011.

As peças policiais selecionadas foram divididas aleatoriamente em dois grupos, formando dois *corpus*, um para validação (teste) dos modelos e outro para treinamento, cada *corpus* composto por 15 documentos.

Para a tarefa de anotação, foi utilizada a ferramenta GATE [JdC07], que possibilita a anotação manual e automática, sendo que para este trabalho foi realizada a anotação manual, a fim de se obter o melhor resultado possível para o treino.

Foram anotadas as entidades pessoa, organização e localização. Diversos foram os desafios enfrentados na tarefa de anotação, dentre eles a ambiguidade na classificação das entidades. Uma rua pode se chamar Rua João da Silva, da mesma forma uma organização pode ter o nome de uma pessoa (por exemplo, Instituição João da Silva).

Para a classe pessoa, foi utilizado o critério de nome próprio, capaz de identificar um ser humano, considerando o nome completo, prenome, sobrenome ou apelido. A classe organização foi definida como qualquer parte do nome que represente uma organização, pública ou privada, com ou sem fins lucrativos. E para localização, foram considerados endereços, nomes de cidades, estados e países.

*Corpus* para validação

O *corpus* construído para validação é composto por um total de 11.296 palavras. Foram anotadas 799 entidades, sendo 374 da classe pessoa, 250 da classe organização e 175 entidades da classe localização.

*Corpus* para treinamento

O *corpus* construído para treinamento é composto por um total de 13.012 palavras. Foram anotadas 806 entidades, sendo 382 da classe pessoa, 278 da classe organização e 146 entidades da classe localização.

### 6.1.2 Modelos

*Convolutional Neural Networks* (CNN)

CNN são Redes Neurais Convolucionais, do inglês *Convolutional Neural Networks*, originalmente criadas para a área de visão computacional, têm alcançado bons resultados na resolução de tarefas de PLN.

As redes convolucionais combinam três ideias arquitetônicas: campos receptivos locais, pesos compartilhados (vinculados) e subamostras espaciais ou temporais.Uma rede neural convolucional consiste em múltiplas partes com funções diferentes.

Inicialmente é comum aplicar sobre o dado de entrada camadas ditas de convolução. Uma camada de convolução é composta por diversos neurônios, cada um responsável por aplicar um filtro em um pedaço específico da imagem. Podemos imaginar cada neurônio sendo conectado a um conjunto de *pixels* da camada anterior e que a cada uma dessa conexões se atribui um peso. A combinação das entradas de um neurônio, utilizando os pesos respectivos de cada uma de suas conexões, produz uma saída passada para a camada seguinte. Os pesos atribuídos às conexões de um neurônio podem ser interpretados como uma matriz que representa o filtro de uma convolução de imagens no domínio espacial (conhecido também como *kernel* ou máscara) [VPV16].

O modelo CNN utilizado neste trabalho foi o português, treinado originalmente no *corpus* Universal Dependencies e WikiNER. Este modelo oferece suporte à identificação das entidades PER (pessoa), LOC (localização), ORG (organização) e MISC (outros).

Para realizar os experimentos com o modelo CNN, foi utilizada a ferramenta Spacy, uma ferramenta de código aberto escrita na linguagem Python que possui 10 modelos de linguagens pré-treinados [Môr18].

LSTM+CNN

Utilizado pelo grupo de PLN da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), treinado no *corpus* LENER-BR [dAdCdO+18], o modelo LSTM+CNN é composto por redes LSTM (*Long Short-Term Memory*) e CNN.

LSTM é um tipo de Rede Neural Recorrente, com uma estrutura computacional complexa que tem obtido sucesso na resolução de tarefas sequenciais. Esta estrutura funciona a partir da ativação de algumas funções, decidindo em alguns momentos se deve manter, alterar ou descartar alguma informação anterior para relacionar com uma informação atual. Dessa forma, leva em consideração um estado passado para relacionar com um estado atual.

## 6.2 Reconhecimento de Entidades Nomeadas

Para a tarefa de Reconhecimento de Entidades Nomeadas (REN), foram realizados dois experimentos, o primeiro utiliza os dois modelos, CNN e LSTM+CNN. Para o segundo experimento, o modelo CNN foi treinado com um *corpus* anotado visando identificar a evolução e desempenho do modelo quando submetido a um treino em um domínio específico.

### 6.2.1 Primeiro experimento

O primeiro experimento teve como resultado um trabalho publicado no STIL-2019 (*Symposium in Information and Human Language Technology*) [MV19], que consiste na aplicação de dois modelos de REN, CNN e LSTM+CNN, na tarefa de análise de documentos policiais.

Para a avaliação, foram selecionadas 15 peças policiais, sendo 05 Autos de Qualificação e Interrogatório, 05 Termos de Declaração e 05 Termos de Depoimento. As peças selecionadas foram produzidas pela Superintendência da Polícia Federal no Rio Grande do Sul, entre os anos de 2010 e 2011, conforme explicado na Seção 2.3.

As classes de entidades selecionadas para o experimento foram pessoa, organização e localização, por serem classes extensivamente estudadas na área de REN, compondo o estado da arte.

Executando os dois modelos, foram obtidos os resultados apresentados na Tabela 6.1 e na Figura 6.1.

Tabela 6.1 – Primeiro experimento: resultado dos modelos

| Modelo | Precisão | Abrangência | Medida-F |
|---|---|---|---|
| CNN | 0,50 | 0,65 | 0,54 |
| LSTM+CNN | 0,70 | 0,69 | 0,68 |

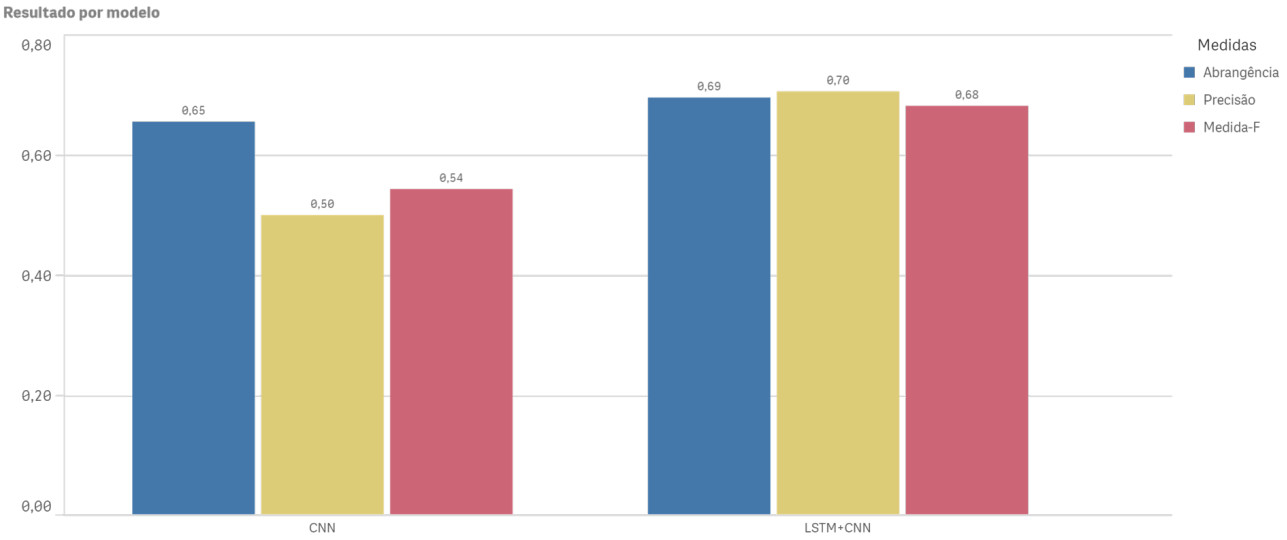


Figura 6.1 – Primeiro experimento: gráfico do resultado

Nota-se que o desempenho foi melhor com a utilização do modelo LSTM+CNN, apresentando uma medida-F de 0,68, contra 0,54 do CNN.

A Figura 6.2 apresenta os resultados em relação às classes de entidades. Verifica-se um desempenho melhor na classe pessoa. Por outro lado, o pior desempenho foi verificado na classe organização.

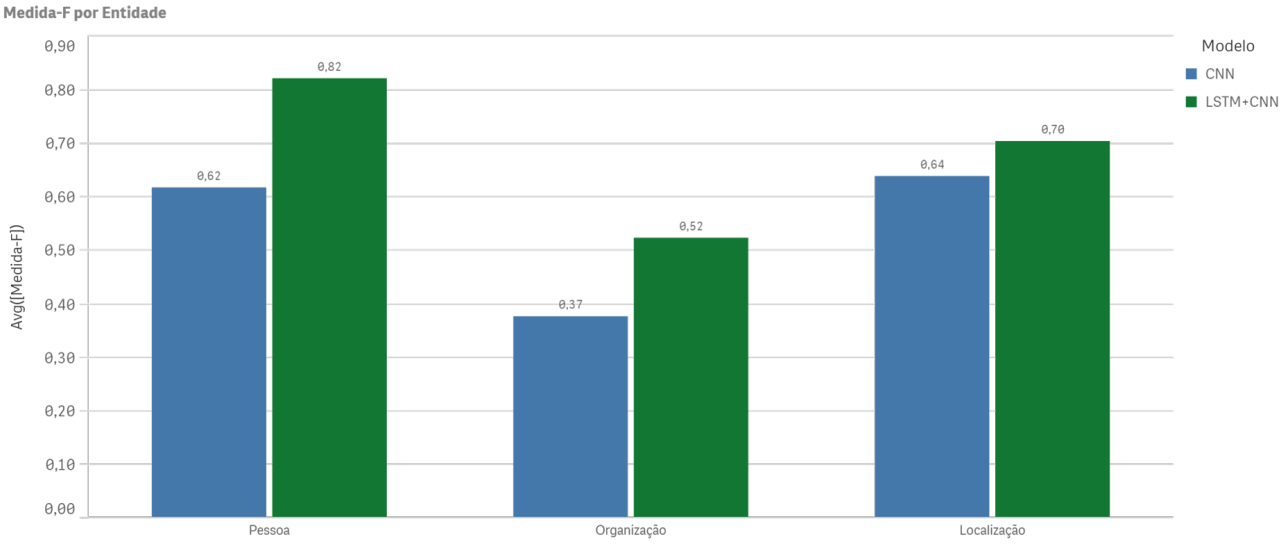


Figura 6.2 – Primeiro experimento: medida-F por classe de entidade

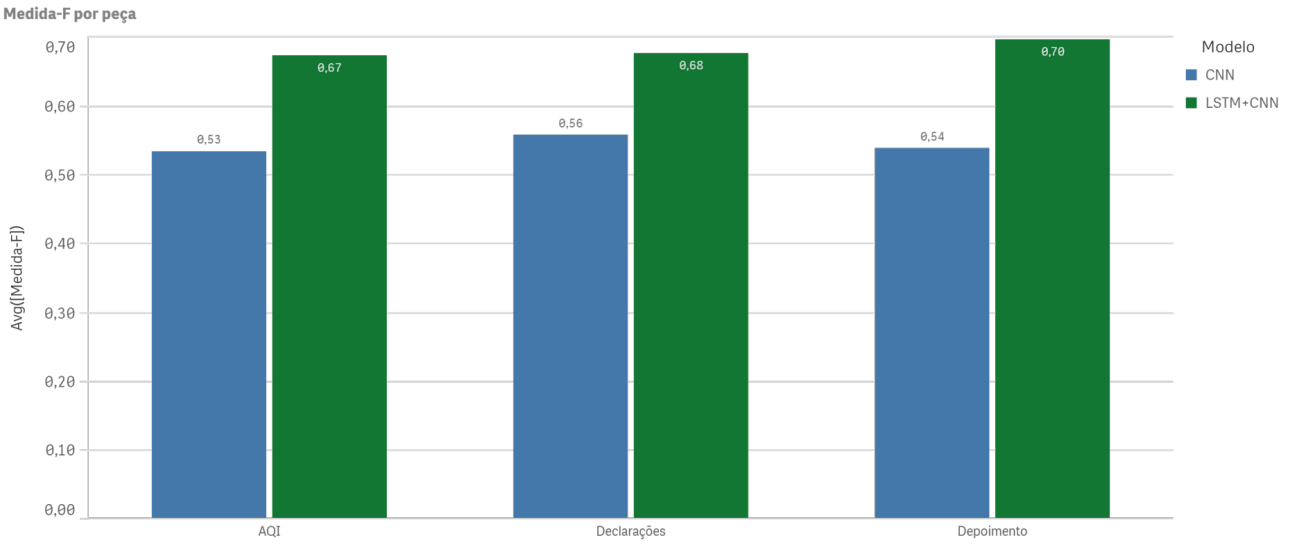


Figura 6.3 – Primeiro experimento: medida-F por peça policial

Na Figura 6.3 são apresentados os resultados separados por peças policiais. Percebe-se que o tipo de documento não altera significativamente o resultado, até mesmo porque os documentos possuem semelhança estrutural.

| | CNN | | | | | | | | | LSTM+CNN | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pessoa | | | Organização | | | Localização | | | Pessoa | | | Organização | | | Localização | | |
| | Prec. | Abr. | M-F | Prec. | Abr. | M-F | Prec. | Abr. | M-F | Prec. | Abr. | M-F | Prec. | Abr. | M-F | Prec. | Abr. | M-F |
| **AQI 01** | 0,64 | 0,58 | 0,61 | 0,30 | 0,38 | 0,33 | 0,50 | 0,75 | 0,60 | 0,91 | 0,83 | 0,87 | 0,50 | 0,25 | 0,33 | 1,00 | 0,42 | 0,59 |
| **AQI 02** | 0,67 | 0,44 | 0,53 | 0,45 | 0,64 | 0,53 | 0,39 | 1,00 | 0,56 | 0,89 | 0,89 | 0,89 | 0,73 | 0,57 | 0,64 | 0,89 | 0,73 | 0,80 |
| **AQI 03** | 0,77 | 0,74 | 0,76 | 0,52 | 0,61 | 0,56 | 0,41 | 0,87 | 0,55 | 0,83 | 0,83 | 0,83 | 0,53 | 0,39 | 0,45 | 0,50 | 0,67 | 0,57 |
| **AQI 04** | 0,63 | 0,45 | 0,53 | 0,27 | 0,33 | 0,30 | 0,58 | 0,92 | 0,71 | 0,90 | 0,82 | 0,86 | 0,50 | 0,67 | 0,57 | 0,89 | 0,67 | 0,76 |
| **AQI 05** | 0,78 | 0,44 | 0,56 | 0,36 | 0,64 | 0,46 | 0,27 | 0,80 | 0,40 | 0,88 | 0,94 | 0,91 | 0,50 | 0,36 | 0,42 | 0,47 | 0,80 | 0,59 |
| **Declarações 01** | 0,50 | 0,33 | 0,40 | 0,44 | 0,44 | 0,44 | 0,55 | 0,86 | 0,67 | 0,75 | 0,67 | 0,71 | 0,67 | 0,44 | 0,53 | 0,67 | 0,57 | 0,62 |
| **Declarações 02** | 0,77 | 0,43 | 0,56 | 0,21 | 0,44 | 0,29 | 0,56 | 1,00 | 0,72 | 0,95 | 0,78 | 0,86 | 0,60 | 0,67 | 0,63 | 0,43 | 0,67 | 0,52 |
| **Declarações 03** | 0,75 | 0,82 | 0,78 | 0,42 | 0,38 | 0,40 | 0,60 | 0,90 | 0,72 | 0,75 | 0,82 | 0,78 | 0,56 | 0,38 | 0,45 | 0,60 | 0,60 | 0,60 |
| **Declarações 04** | 0,75 | 0,38 | 0,50 | 0,44 | 0,53 | 0,48 | 0,36 | 0,89 | 0,52 | 0,86 | 0,75 | 0,80 | 0,64 | 0,60 | 0,62 | 0,75 | 0,67 | 0,71 |
| **Declarações 05** | 0,73 | 0,69 | 0,71 | 0,33 | 0,45 | 0,38 | 0,64 | 1,00 | 0,78 | 0,94 | 1,00 | 0,97 | 0,83 | 0,45 | 0,59 | 0,77 | 0,71 | 0,74 |
| **Depoimento 01** | 0,67 | 0,75 | 0,71 | 0,00 | 0,00 | 0,00 | 0,50 | 1,00 | 0,67 | 0,71 | 0,63 | 0,67 | 0,14 | 1,00 | 0,25 | 0,86 | 1,00 | 0,92 |
| **Depoimento 02** | 0,53 | 0,82 | 0,64 | 0,25 | 0,25 | 0,25 | 0,28 | 0,88 | 0,42 | 0,90 | 0,82 | 0,86 | 0,57 | 0,75 | 0,65 | 0,70 | 0,88 | 0,78 |
| **Depoimento 03** | 0,55 | 0,75 | 0,63 | 0,27 | 0,44 | 0,33 | 0,53 | 1,00 | 0,69 | 0,78 | 0,88 | 0,82 | 0,40 | 0,44 | 0,42 | 0,73 | 0,89 | 0,80 |
| **Depoimento 04** | 0,50 | 1,00 | 0,67 | 0,56 | 0,56 | 0,56 | 0,67 | 1,00 | 0,80 | 1,00 | 0,80 | 0,89 | 0,56 | 0,56 | 0,56 | 0,82 | 0,90 | 0,86 |
| **Depoimento 05** | 0,58 | 0,70 | 0,64 | 0,33 | 0,29 | 0,31 | 0,64 | 0,88 | 0,74 | 0,55 | 0,60 | 0,57 | 0,60 | 0,86 | 0,71 | 0,71 | 0,63 | 0,67 |

Figura 6.4 – Primeiro experimento: resultado geral

Na Figura 6.4 são apresentados os resultados de todo o experimento. Nota-se a distribuição dos resultados, evidenciando melhores resultados para o modelo LSTM+CNN na tarefa de REN para as classes pessoa, localização e organização.

### 6.2.2 Segundo experimento

O objetivo deste segundo experimento é a construção de um *corpus* anotado para treino, visando analisar sua contribuição para a tarefa proposta.

Optou-se por treinar apenas um dos modelos em razão dos recursos e tempo disponíveis, uma vez que o objetivo não é o de avaliar o melhor modelo, mas analisar a contribuição que o treino com um *corpus* anotado pode oferecer, e os benefícios à área de investigação de crimes de lavagem de dinheiro. Sendo assim, o modelo treinado foi o CNN por ter apresentado um resultado inferior no primeiro experimento, proporcionando uma melhor análise de sua evolução.

#### Treino do modelo CNN

Com o *corpus* devidamente anotado (Seção 6.1.1), foi realizado o treino do modelo CNN a fim de analisar sua evolução.

O método utilizado para treino está apresentado no Apêndice A. Utiliza-se um modelo inicial como referência, neste caso foi utilizado o modelo padrão da língua portuguesa (pt_core_news_sm), treinado a partir de textos jornalísticos da CNN e da WikiNER.

#### Resultados

Para avaliação do modelo CNN após o treino, foram utilizadas as mesmas peças do primeiro experimento e comparados os resultados. Os valores do modelo CNN após treino passam a ser rotulados como CNN(T).

Analisando a Tabela 6.2, verifica-se que os resultados do modelo CNN, antes do treino, estavam todos abaixo do LSTM+CNN. Após o treino, percebe-se uma significativa melhora nos resultados do CNN(T), alcançando uma medida-F de 0,78 (0,24 maior que o resultado anterior).

Não se trata de uma comparação entre os modelos, até porque o LSTM+CNN não foi treinado, contudo, trata-se de uma constatação do quanto o treino de um modelo de REN pode melhorar seu desempenho quando treinado a partir de um *corpus* anotado com documentos pertencentes ao domínio analisado.

Tabela 6.2 – Segundo experimento: resultado dos modelos

| Modelo | Precisão | Abrangência | Medida-F |
|---|---|---|---|
| CNN | 0,50 | 0,65 | 0,54 |
| CNN(T) | 0,83 | 0,77 | 0,78 |
| LSTM+CNN | 0,70 | 0,69 | 0,68 |

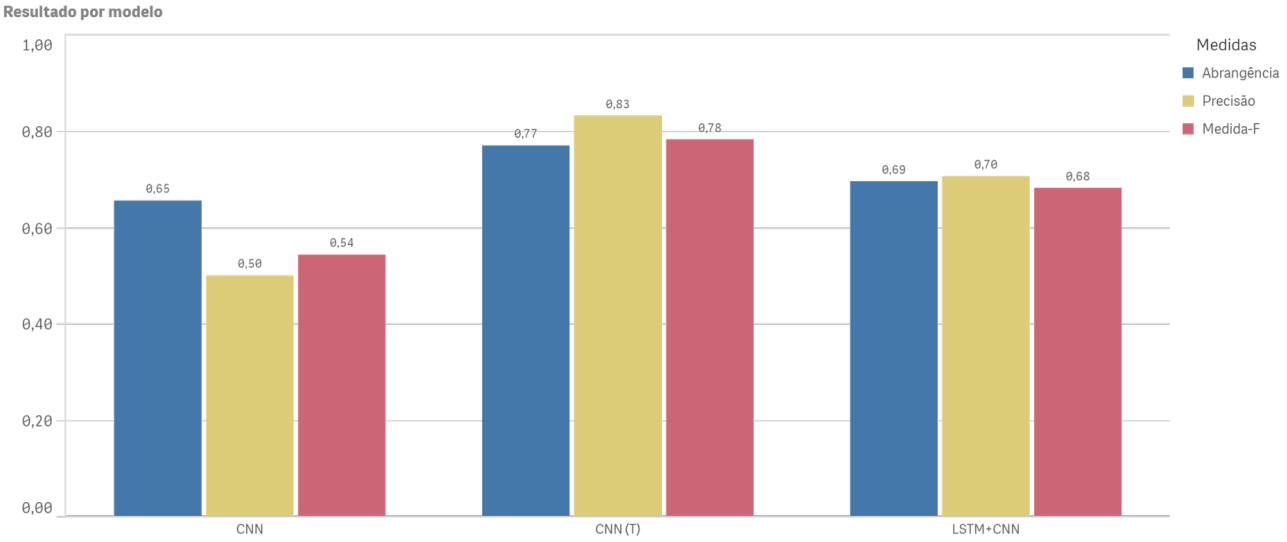


Figura 6.5 – Segundo experimento: gráfico do resultado

Analisando o gráfico do resultado (Figura 6.5), observa-se uma evolução em todas as medidas do modelo CNN(T), sendo que, comparando com os resultados antes do treino, houve uma evolução de 0,12 pontos na abrangência, 0,33 pontos na precisão, e 0,24 pontos na medida-F.

Após o treino, o modelo CNN(T) obteve melhores resultados em todas as classes de entidades, como se observa no gráfico da Figura 6.6. Vale ressaltar a evolução obtida na classe organização, que no primeiro experimento foi o pior resultado, em razão da ambiguidade e dificuldade em classificar uma entidade como pessoa ou organização.

Verifica-se, portanto, que o modelo CNN(T) obteve um resultado de 0,39 pontos superior ao obtido antes do treino, superando até mesmo o resultado obtido na classe localização.

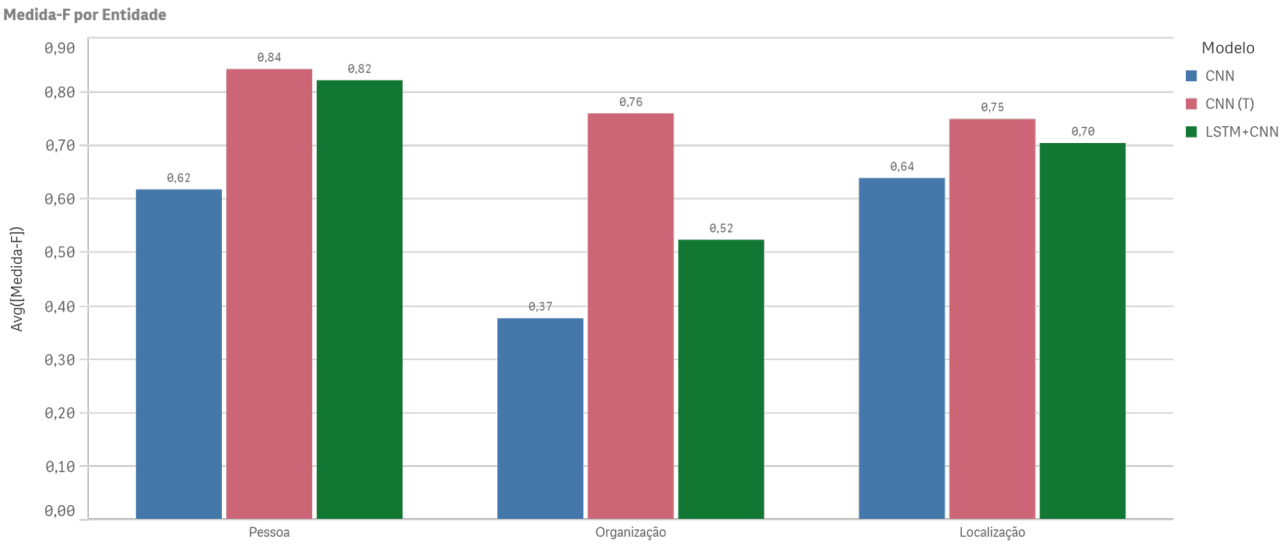


Figura 6.6 – Segundo experimento: medida-F por classe de entidade

Analisando o gráfico por peça policial (Figura 6.7), note-se a melhora substancial nos resultados do modelo CNN(T), e observa-se, como já constatado no primeiro experimento (Figura 6.3), que os resultados são semelhantes em todas as três dimensões (AQI, Declarações e Depoimento), levando a crer que a escolha do tipo de peça policial não é determinante para o desempenho dos modelos neste experimento.

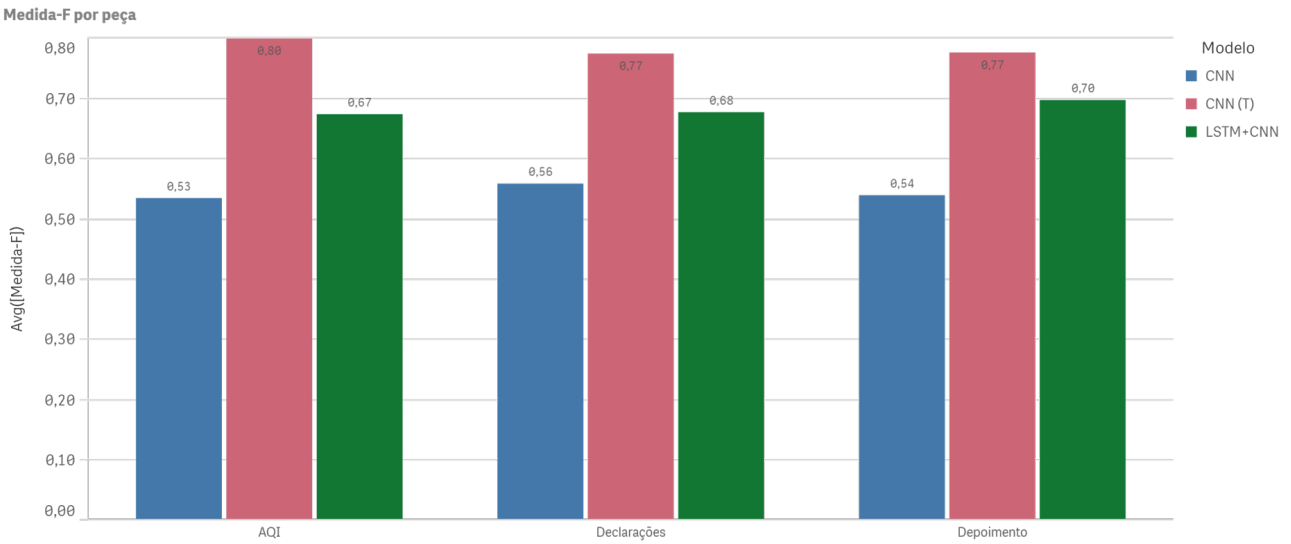


Figura 6.7 – Segundo experimento: medida-F por peça policial

A seguir, são apresentados gráficos do tipo *boxplot* (Figuras 6.8, 6.9 e 6.10). O *boxplot* é uma maneira gráfica de representar a alteração dos dados de uma variável por meio de quartis e a distribuição dos dados.

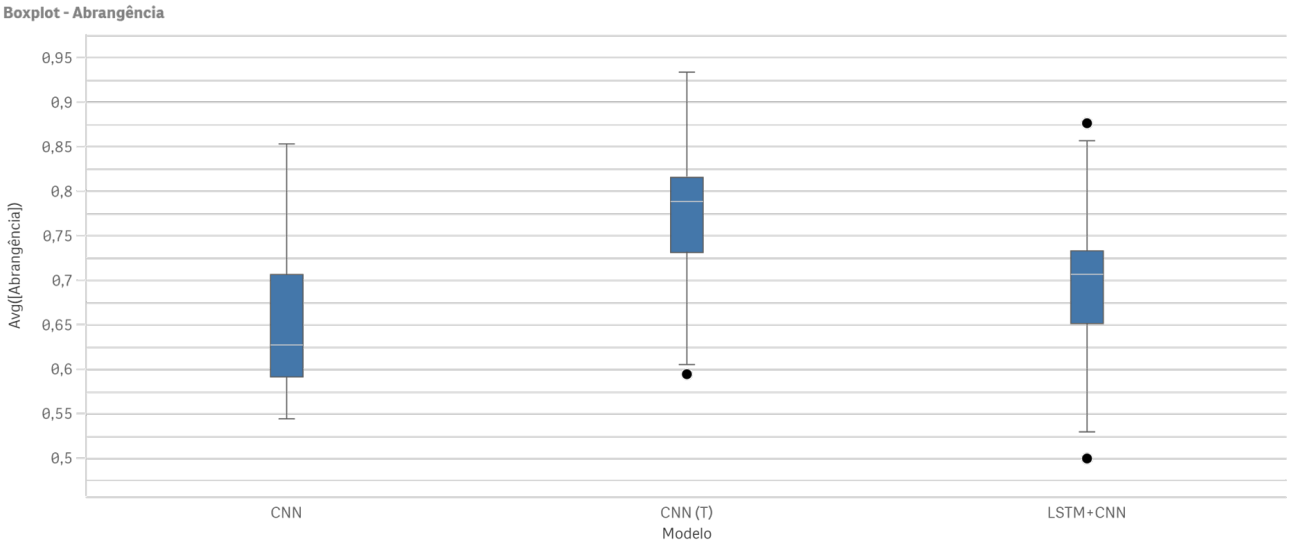


Figura 6.8 – Segundo experimento: *boxplot* abrangência

Neste tipo de representação, é possível, ainda, observar os valores discrepantes (*outliers*), no gráfico representados pelos círculos na cor preta, que são aqueles valores muito diferentes do restante do conjunto de dados [RR02].

Dentre as possibilidades de análise dos resultados, podemos observar a variação de cada modelo, representada pelo distanciamento no eixo y, bem como a amplitude de cada *box*, representando a concentração dos resultados. Quanto menor a amplitude, mais consistente tende a ser o modelo.

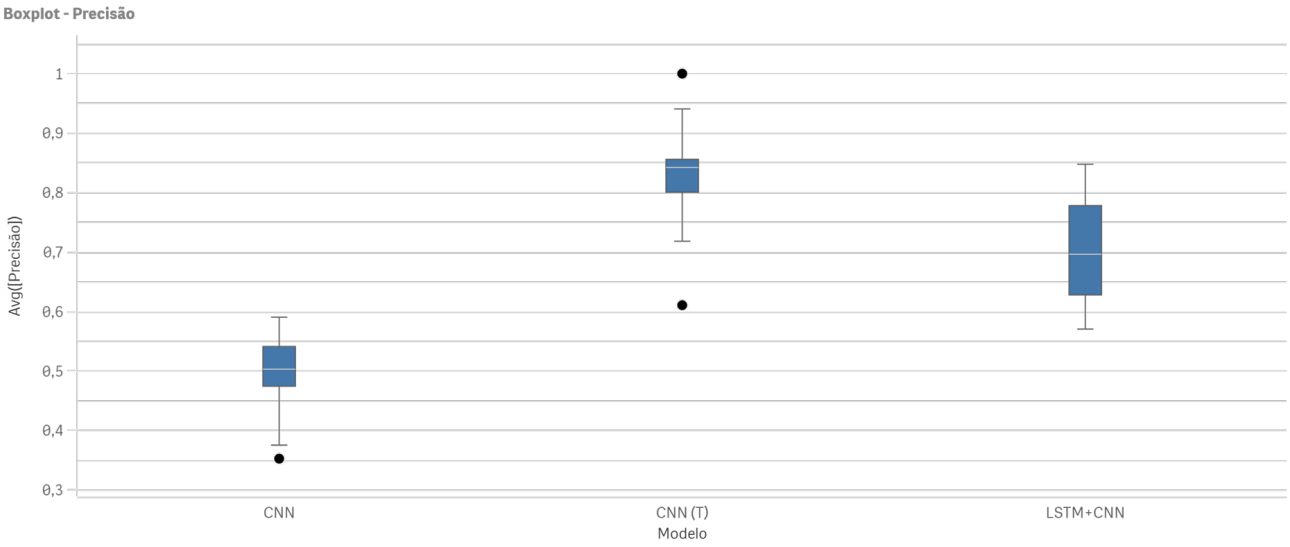


Figura 6.9 – Segundo experimento: *boxplot* precisão

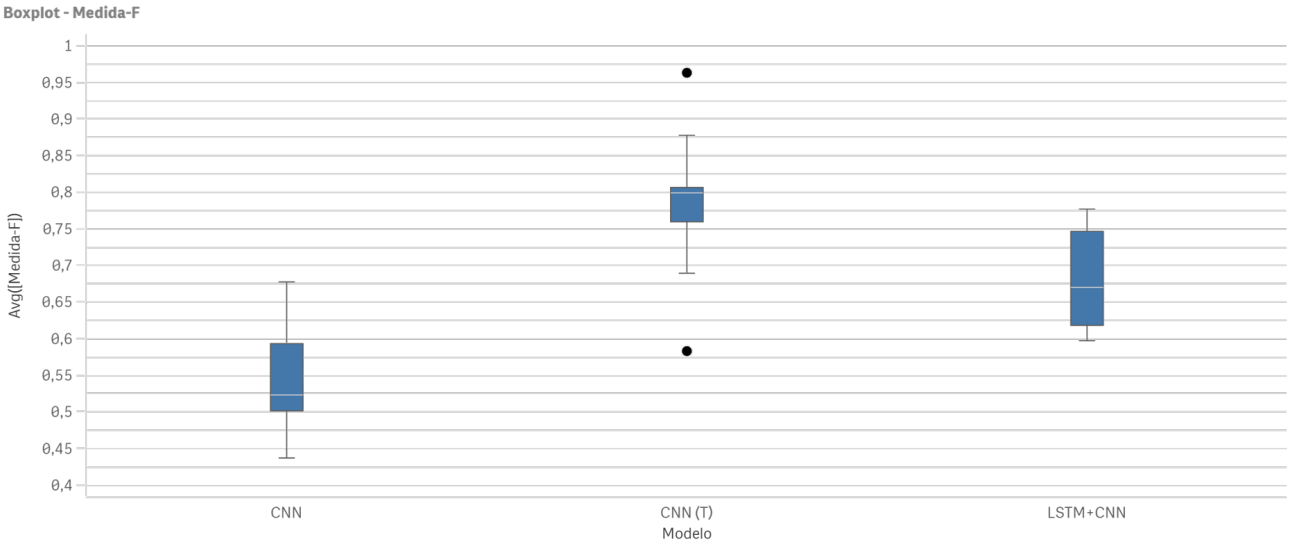


Figura 6.10 – Segundo experimento: *boxplot* medida-F

Observa-se, portanto, após análise dos gráficos de *boxplot*, que houve uma melhora significativa nos resultados do modelo CNN(T), bem como a redução da amplitude, o que representa uma maior consistência na tarefa de REN.

A Figura 6.11 apresenta o gráfico de dispersão com a média da precisão no eixo Y, da abrangência no eixo X e o tamanho do círculo representa a média da medida-F.

A análise deste gráfico possibilita visualizar o distanciamento do modelo CNN(T) para CNN em todos os aspectos, demonstrando a evolução dos resultados a partir das técnicas utilizadas neste segundo experimento.

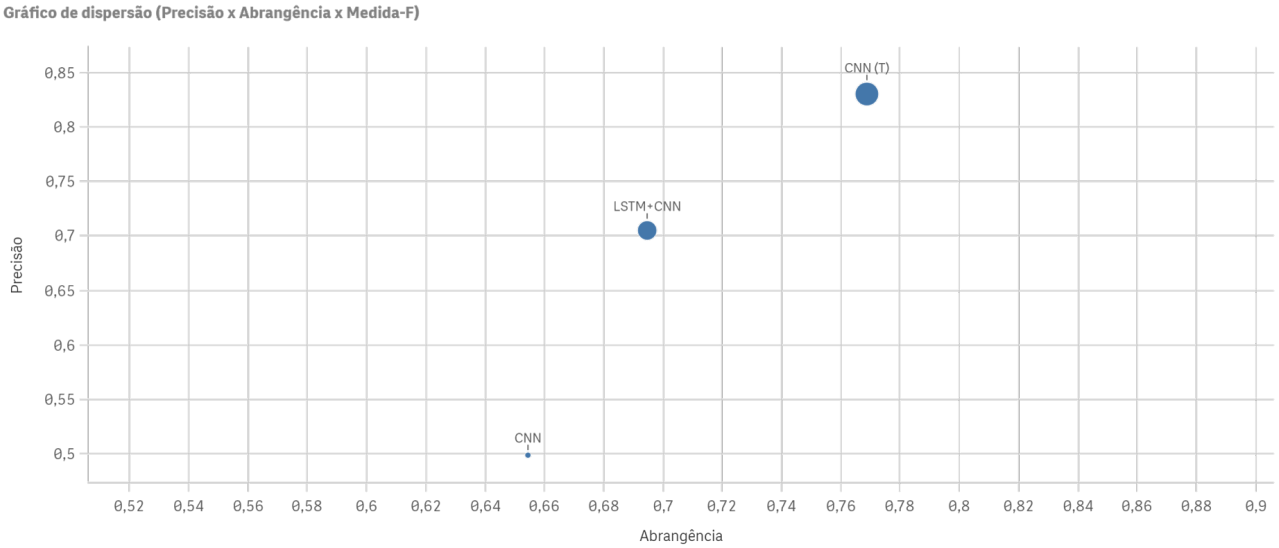


Figura 6.11 – Segundo experimento: gráfico de dispersão

Análise de erros e desafios

Nesta subseção são apresentados alguns erros e desafios identificados nos experimentos realizados.

O primeiro experimento consistiu em processar o *corpus* selecionado nos modelos de REN com a configuração padrão, exigindo um estudo prévio dos conceitos e dos modelos utilizadas. Os primeiros resultados já apresentaram a dificuldade em classificar corretamente a classe de entidade organização. Notou-se que, o fato das pessoas e das organizações serem padronizadas com a escrita em caixa alta e ocuparem frequentemente a função sintática de sujeito, gera, frequentemente, a ocorrência de ambiguidade.

Com o objetivo de melhorar os resultados, foi selecionado um novo *corpus* para treino e, inicialmente, foi realizada a anotação apenas da classe pessoa. Essa primeira anotação não demonstrou ser suficiente, já que a ambiguidade envolvia mais de uma classe, principalmente pessoa e organização. O *corpus* de treino foi então anotado com as três classes envolvidas no experimento.

A anotação, por sua vez, demonstrou ser uma tarefa complexa, envolvendo questões de ambiguidades também enfrentadas pelos especialistas, necessitando de critérios pré-estabelecidos.

Os resultados pós-treino apresentaram uma melhora significativa na classificação, demonstrando a relevância do treino do modelo no domínio específico em que se pretende aplicar o REN.

Conclusão dos experimentos

Analisando os resultados apresentados neste capítulo, verifica-se a relevância da utilização de um *corpus* anotado como forma de treinar modelos de REN, uma vez que o vocabulário e a sintaxe são característicos em cada domínio. Nos experimentos realizados, a evolução do modelo CNN treinado foi de 66% na precisão, 18% na abrangência e 44% na medida-F.

Constata-se que a tarefa de Reconhecimento de Entidades Nomeadas em documentos policiais é uma técnica de PLN aplicável e eficiente para as investigações, uma vez que um dos principais objetivos de uma investigação policial é a identificação da autoria delitiva (Seção 2.1). Dado o grande volume de informações envolvidas em uma investigação, bem como a variedade e complexidade dos documentos analisados, os modelos de REN podem proporcionar agilidade e efetividade na identificação da autoria, assim como na elucidação de crimes.

## 6.3 Outras técnicas de PLN

Outras técnicas de Processamento de Linguagem Natural podem ser utilizadas nas investigações policiais e serão a seguir apresentadas como forma de expandir as possibilidades aplicáveis. O *dataset* utilizado foi o mesmo dos experimentos anteriores.

### 6.3.1 Pré-processamento do texto

O pré-processamento do texto é uma importante etapa para a preparação dos dados (Seção 3.4). A primeira etapa é a transformação, o objetivo é retirar características das palavras que não alteram seu significado e auxiliam na análise. As técnicas aplicadas foram as de *lowercase* e remoção de acentos, a primeira evita que palavras com o mesmo significado iniciadas com caracteres em maiúsculo venham a ser diferenciadas de uma palavra semelhante iniciada com caracteres em minúsculo, a segunda técnica remove acentos, corrigindo distinções ocasionadas por erro de digitação.

Na segunda etapa, foi utilizada a técnica de tokenização (Seção 3.4.1) para a separação das palavras delimitadas por espaços.

Em seguida, foram aplicados filtros com a utilização de expressões regulares (regex) para a remoção de caracteres especiais, além da remoção de *stopwords* com uma lista de 202 palavras, dentre elas artigos, pronomes, preposições e conjunções. A lista de *stopwords* foi criada com base nas palavras mais utilizadas nos modelos de peças policiais apresentados na Seção 2.3.

### 6.3.2 *Bag-of-words*

Apesar de ser um modelo relativamente básico, o *bag-of-words* é frequentemente usado para tarefas de Processamento de Linguagem Natural, como classificação de texto. Para facilitar a manipulação e análise dos dados, utiliza-se nesta fase uma técnica de vetorização, transformação do texto em vetores, computando a frequência com que uma palavra aparece.

Foram configurados três parâmetros nesta tarefa: *term frequency*, *document frequency* e *regularization*. Para o *term frequency* foi utilizado o valor "*count*" que conta o número de ocorrência da palavra no documento. Para o parâmetro *document frequency* foi selecionado o IDF (*Inverse Document Frequency*), uma medida estatística que indica a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um *corpus* linguístico. Ela é frequentemente utilizada como fator de ponderação na recuperação de informações e na mineração de dados. O valor de uma palavra aumenta proporcionalmente à medida que aumenta o número de ocorrências dela em um documento, no entanto, esse valor é equilibrado pela frequência da palavra no *corpus*. Isso auxilia a distinguir o fato da ocorrência de algumas palavras serem geralmente mais comuns que outras. E no parâmetro *regularization* foi atribuído o valor de L2 (Euclidean) que normaliza o comprimento do vetor para a soma dos quadrados.

### 6.3.3 Nuvem de palavras

Com os dados pré-processados e transformados em vetor, já é possível utilizar os resultados para análise e produção de conhecimento, como é o caso da visualização dos dados pela técnica de nuvem de palavras.

Utilizando-se da frequência com que cada Entidade Nomeada (EN) da classe pessoa ocorre nos documentos, obteve-se a nuvem de palavras representada na Figura 6.12.


Figura 6.12 – Nuvem de palavras para a classe pessoa (dados anonimizados)

Observa-se que os nomes foram anonimizados em razão do Termo de Confidencialidade firmado com a Polícia Federal, visando a proteção das informações pessoais e das empresas constantes nos documentos.

A utilização dessa técnica pelo policial analista proporciona uma visão macro das pessoas envolvidas na investigação e facilita a obtenção de *insights*. Esta técnica também pode ser utilizada para outras classes de EN ou grupos de palavras.

### 6.3.4 Escala Multidimensional (MDS)

Utilizando mais uma técnica de visualização, a Escala Multidimensional (MDS) provém de uma família de técnicas de análise de proximidade de dados e tem se mostrado um importante instrumento matemático de mensuração [SBBM09].

A técnica consiste em comparar os objetos em vários parâmetros concomitantemente, não se trata de fazer uma avaliação com um parâmetro de cada vez. A proximidade encontrada entre os objetos refletirá o grau de similaridade e será obtida a partir da distância entre eles, representados graficamente por meio de pontos em um espaço euclidiano, como representado na Figura 6.13.

Novamente os nomes foram anonimizados em razão do Termo de Confidencialidade firmado com a Polícia Federal, visando a proteção das informações pessoais e das empresas constantes nos documentos.

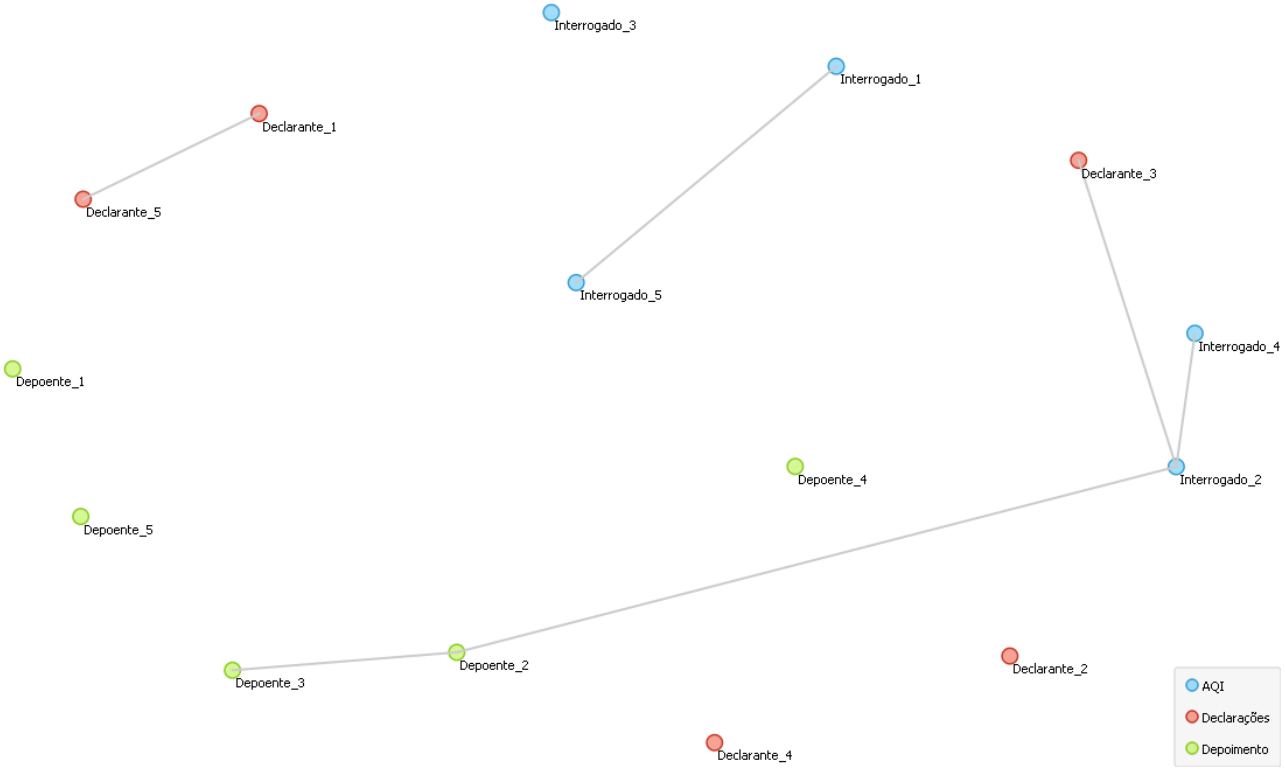


Figura 6.13 – Escala Multidimensional (MDS), dados anonimizados

Na representação acima, cada documento é identificado com uma cor, de acordo com o tipo de documento (peça policial). O documento é etiquetado com o nome da pessoa que prestou o depoimento, a declaração ou o AQI. Sendo assim, em uma grande coleção de documentos, por meio da ligação entre eles, é possível identificar quais documentos possuem similaridade.

Na Figura 6.13, observa-se que existe similaridade entre os depoimentos do Depoente_2 e Depoente_3, a declaração do Declarante_3 e os AQIs dos interrogados Interrogado_2 e Interrogado_4. Com o auxílio de um especialista, verificou-se que, em que pese as peças acima relacionadas não pertençam ao mesmo inquérito policial, elas possuem similaridade quanto às pessoas e organizações investigadas.

Verifica-se que a visualização da similaridade por meio da Escala Multidimensional proporciona ao investigador uma poderosa ferramenta de análise. Por meio desta técnica, o policial poderá otimizar sua análise agrupando os diversos documentos conforme sua similaridade, agilizando o trabalho de análise de grandes coleções de documentos.

### 6.3.5 Comparação com a pesquisa

Após a utilização das técnicas de PLN apresentadas neste capítulo, foi possível comparar os resultados obtidos às informações relevantes da pesquisa realizada com especialistas da Polícia Federal apresentada no Capítulo 5.

Das 15 classes de EN apresentadas na pesquisa, 3 foram abordadas neste trabalho (pessoa, organização e localização).

Dos 19 termos que indicam relações entre entidades identificados na pesquisa, 13 foram verificados no *dataset* deste trabalho: socio, proprietario, procurador, procuracao, servidor, funcionario, empregado, indicado, parceiro, auxiliar, conhece, desconhece e ouviu. As palavras estão sem acentuação em razão do pré-processamento realizado.

Os verbos identificados no *dataset* foram: mandar, determinar, auferir, lucrar, adquirir, transferir, remeter, pedir, solicitar, exigir, receber, comprar, enganar, assinar, usar, vender, prejudicar, doar, intermediar, ceder e alugar. Para a identificação dos verbos, foi utilizada a técnica de normalização do texto com a utilização do algoritmo *Porter Stemming*.

As relações entre entidades e os verbos identificados não foram explorados neste trabalho, mas ficam como sugestão para futuros trabalhos (Seção 7.2), sendo úteis para o desenvolvimento de tarefas de PLN que envolvam Extração de Relação entre Entidades Nomeadas (RE) e análises sintáticas.

# 7. CONSIDERAÇÕES FINAIS

Um dos grandes desafios da justiça brasileira, em especial a criminal, é a celeridade dos processos. Além das questões processuais que envolvem diversas fases recursais, existe uma deficiência estrutural que, dentre outros problemas, geralmente está relacionada à tecnologia. A polícia judiciária, a exemplo da Polícia Federal, possui um relevante papel na persecução penal, sendo fundamental na identificação da autoria e materialidade criminal.

Em razão da complexidade e do volume crescente de dados e documentos presentes nas investigações de crimes de financeiros e de lavagem de dinheiro, verifica-se a necessidade do uso da tecnologia no apoio à análise. A partir de um estudo de caso da Polícia Federal brasileira, foram apresentadas soluções tecnológicas para a análise de documentos produzidos e apreendidos nas investigações policiais.

Inicialmente, foi realizada uma pesquisa com especialistas da Polícia Federal que atuam nas investigações de crimes financeiros, visando identificar as principais informações que se pretende obter ao analisar os documentos numa investigação policial.

Para a realização dos experimentos, foi utilizado um *corpus* fornecido pela Superintendência da Polícia Federal no Rio Grande do Sul. Este *corpus* é composto por peças policiais produzidas a partir de depoimentos e interrogatórios realizados pela Polícia Federal entre os anos de 2010 e 2011.

A partir das informações obtidas na pesquisa e do *corpus* fornecido, foram realizados experimentos utilizando técnicas de Processamento de Linguagem Natural (PLN) e os resultados foram apresentados e analisados.

Foram realizados dois experimentos para a tarefa de Reconhecimento de Entidades Nomeadas (REN). No primeiro experimento foram utilizados dois modelos (CNN e LSTM+CNN), o modelo LSTM+CNN demonstrou maior eficiência frente ao CNN, superando os resultados nas três classes de entidades e nos três tipos de documentos propostos para análise.

Visando aumentar o desempenho dos modelos de PLN para extração de informações em textos livres, o segundo experimento consistiu em treinar o modelo que obteve o pior resultado (CNN) com documentos do mesmo domínio analisado, a partir de um novo *corpus* contendo anotação das classe pessoa, localização e organização. Os resultados demonstraram que o treino melhorou significativamente o desempenho do modelo, diminuindo boa parte da ambiguidade e melhorando a precisão.

Constata-se que, apesar de não ser uma tarefa simples, treinar um modelo de REN com um *corpus* anotado é relevante para um bom desempenho, considerando a variedade de tipos de documentos a serem analisados em uma investigação policial.

Outras técnicas de PLN também foram estudadas e aplicadas neste trabalho como forma de demonstrar sua utilidade e relevância na descoberta do conhecimento em dados não estruturados. A utilização de técnicas de similaridade entre documentos, a exemplo da MDS, auxiliam na seleção e estruturação dos documentos a serem analisados. As técnicas de visualização, como as nuvens de palavras, oferecem recursos visuais que facilitam a análise dos dados, proporcionando celeridade e eficiência à análise.

Dessa forma, conclui-se que, dado o grande volume de dados não estruturados a serem analisados em uma investigação policial, as técnicas e modelos aqui apresentados oferecem um importante suporte no trabalho investigativo, auxiliando na identificação de pessoas e de elementos que servirão para o seguimento da persecução penal.

## 7.1 Limitações

Optou-se por treinar apenas um dos modelos (CNN) em razão dos recursos e tempo disponíveis, uma vez que o objetivo não foi avaliar o melhor modelo, mas analisar a contribuição que eles podem oferecer à área de investigação de crimes de lavagem de dinheiro.

Outra limitação diz respeito ao *corpus* utilizado neste trabalho, selecionado a partir de peças policiais produzidas em cartório que, mesmo sendo compostas em sua maior parte de texto livre, possuem uma estrutura padronizada. Isso faz com que a tarefa de PLN seja facilitada, se comparada à utilização de um *corpus* aleatório, que contém diversos tipos de documentos (planilhas, extratos bancários, procurações, escrituras, etc.).

As Entidades Nomeadas foram limitadas às classes pessoa, localização e organização, por serem as EN mais utilizadas na literatura e por consistirem em padrões utilizados nos modelos na língua portuguesa.

## 7.2 Trabalhos futuros

Como trabalhos futuros, sugerem-se:

- Treinar o modelo LSTM+CNN, conforme o segundo experimento (Seção 6.2.2), para comparar seu desempenho ao CNN;

- Acrescentar as demais classes de EN identificadas na pesquisa realizada com especialistas (Capítulo 5);

- Visando aprimorar a tarefa de REN aplicada à investigação, sugere-se a anotação de um *corpus* maior e diversificado, contendo mais tipos de documentos, submetendo os modelos ao treino e nova avaliação;
- Em complemento à tarefa de REN, utilizando os termos e verbos identificados na pesquisa, sugere-se a Extração das Relações entre as Entidades Nomeadas (RE), explorando o elo entre as entidades identificadas; e
- Desenvolvimento de uma ferramenta que aplique os conhecimentos expostos neste trabalho, gerando soluções para as instituições de polícia judiciária, em especial a Polícia Federal.

# REFERÊNCIAS BIBLIOGRÁFICAS


[Bad16] Badaró, G. “Lavagem de dinheiro: o conceito de produto indireto da infração penal antecedente no crime de lavagem de dinheiro”, *Revista dos Tribunais*, vol. 967, Maio 2016, pp. 73–93.

[Car04] Cardoso, O. N. P. “Recuperação de informação.”, *Journal of Computer Science*, vol. 2, Novembro 2004, pp. 33–38.

[Cas17] Casseb, Á. M. “A necessária aplicação de técnicas inovadoras de análise financeira pelos órgãos de persecução criminal nas investigações e no combate ao crime de lavagem de capitais”, Monografia, Instituto Brasiliense de Direito Público, 2017, 35p.

[CMV16] Collovini, S.; Machado, G.; Vieira, R. “Extracting and structuring open relations from portuguese text”. In: International Conference on Computational Processing of the Portuguese Language, 2016, pp. 153–164.

[CNC+19] Collovini, S.; Neto, J. F. S.; Consoli, B. S.; Terra, J.; Vieira, R.; Quaresma, P.; Souza, M.; Claro, D. B.; Glauber, R. “Portuguese named entity recognition and relation extraction tasks”. In: Iberian Languages Evaluation Forum, 2019, pp. 390–410.

[dAdCdO+18] de Araujo, P. H. L.; de Campos, T. E.; de Oliveira, R. R.; Stauffer, M.; Couto, S.; Bermejo, P. “Lener-br: a dataset for named entity recognition in brazilian legal text”. In: International Conference on Computational Processing of the Portuguese Language, 2018, pp. 313–323.

[dAV14] do Amaral, D. O. F.; Vieira, R. “Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields.”, *Linguamática*, vol. 6, Julho 2014, pp. 41–49.

[dCdSS19] de Castro, P. V. Q.; da Silva, N. F. F.; Soares, A. d. S. “Contextual representations and semi-supervised named entity recognition for portuguese language”. In: Iberian Languages Evaluation Forum, 2019, pp. 411–420.

[DE 11] DE CARLI, C. V. “Lavagem de dinheiro: prevenção e controle penal”. Verbo Jurídico, 2011, 782p.

[dLJ07] de LEMOS JÚNIOR, A. P. “Uma reflexão sobre as dificuidades da investigação criminal do crime de lavagem de dinheiro”, *Revista Justitia*, vol. 197, Julho-Dezembro 2007, pp. 23–35.

[FCLC01] Freitas, C. M. D. S.; Chubachi, O. M.; Luzzardi, P. R. G.; Cava, R. A. “Introdução à visualização de informações”, *Revista de Informática Teórica e Aplicada*, vol. 8, Outubro 2001, pp. 143–158.

[FJD07] Ferro Júnior, C. M.; Dantas, G. F. d. L. “A descoberta e a análise de vínculos na complexidade da investigação criminal moderna”. Capturado em: https://jus.com.br/artigos/10002, Janeiro 2020.

[Fox89] Fox, C. “A stop list for general text”. In: Special Interest Group on Information Retrieval Forum, 1989, pp. 19–21.

[FS07] Feldens, L.; Schmidt, A. Z. “Investigação criminal e ação penal”. Livraria do Advogado, 2007, 169p.

[Gom95] Gomes, L. F. “A impunidade da macro delinquência econômica desde a perspectiva criminológica da teoria da aprendizagem”, *Revista Brasileira de Ciências Criminais*, vol. 3, Julho-Setembro 1995, pp. 166–174.

[HLLE14] Heimerl, F.; Lohmann, S.; Lange, S.; Ertl, T. “Word cloud explorer: text analytics based on word clouds”. In: Proceedings of the 47th Hawaii International Conference on System Sciences, 2014, pp. 1833–1842.

[Ima01] Imamura, C. Y.-M. “Pré-processamento para extraçao de conhecimento de bases textuais”, Dissertação de Mestrado, Universidade de São Paulo, 2001, 116p.

[JC11] Junior, O. D.; Claro, D. B. “Uma análise do reconhecimento textual de nomes de pessoas e organizações na computação forense”. In: Proceeding of the 6th International Conference on Forensic Computer Science, 2011, pp. 7–15.

[JdC07] Junior, E. A.; de Carvalho, C. L. “Processamento de linguagens naturais e a ferramenta gate”, Relatório Técnico, Instituto de Informática da Universidade Federal de Goiás, 2007, 24p.

[KF67] Kučera, H.; Francis, W. N. “Computational analysis of present-day american english”. Dartmouth Publishing Group, 1967, 424p.

[KIL08] Ku, C. H.; Iriberri, A.; Leroy, G. “Natural language processing and e-government: crime information extraction from heterogeneous data sources”. In: Proceedings of the International Conference on Digital Government Research, 2008, pp. 162–170.

[KL13] Ku, C.-H.; Leroy, G. “Automated crime report analysis and classification for e-government and decision support”. In: Proceedings of the 14th Annual International Conference on Digital Government Research, 2013, pp. 18–27.

[Mir03] Mirabete, J. F. “Processo penal”. Atlas, 2003, 835p.

[Mis11] Misse, M. “O papel do inquérito policial no processo de incriminação no brasil: algumas reflexões a partir de uma pesquisa”, *Sociedade e Estado*, vol. 26, Abril 2011, pp. 15–27.

[Moh14] Mohit, B. “Named entity recognition”. In: *Natural Language Processing of Semitic Languages*, Springer, 2014, pp. 221–245.

[Môr18] Môro, D. K. “Reconhecimento de entidades nomeadas em documentos de língua portuguesa”, Monografia, Universidade Federal de Santa Catarina, 2018, 38p.

[MV19] Moreira, F.; Vieira, R. “Aplicação de reconhecimento de entidades nomeadas em investigação de crimes financeiros”. In: Proceedings of the 2nd Symposium in Information and Human Language Technology, 2019, pp. 134–143.

[MZR16] Machado, B. A.; Zackseski, C.; Raupp, R. M. “A investigação e a persecução penal da corrupção e dos delitos econômicos: uma análise exploratória do sistema de justiça federal”, *Revista Brasileira de Ciências Criminais*, vol. 118, Janeiro 2016, pp. 299–329.

[Nov10] Nova Júnior, H. d. A. S. V. “Visualização de informação como ferramenta de auxílio na avaliação formativa em educação a distância”, Dissertação de Mestrado, Universidade Federal de Pernambuco, 2010, 110p.

[NS07] Nadeau, D.; Sekine, S. “A survey of named entity recognition and classification”, *Lingvisticae Investigationes*, vol. 30, Janeiro 2007, pp. 3–26.

[PAS+19] Pirovani, J.; Alves, J.; Spalenza, M.; Silva, W.; da Silveira Colombo, C.; Oliveira, E. “Adapting ner (crf+ lg) for many textual genres”. In: Proceedings of the 35th Conference of the Spanish Society for Natural Language Processing, 2019, pp. 421–433.

[Pau08] Paulovich, F. V. “Mapeamento de dados multi-dimensionais - integrando mineração e visualização”, Tese de Doutorado, Universidade de São Paulo, 2008, 144p.

[Pir15] Pires, J. C. B. “Extração e mineração de informação independente de domínios da web na língua portuguesa”, Dissertação de Mestrado, Universidade Federal de Goiás, 2015, 93p.

[PML96] Percy, C. E.; Meyer, C. F.; Lancashire, I. “Synchronic corpus linguistics: papers from the sixteenth international conference on english language research on computerized corpora”. Rodopi Publishers, 1996, 289p.

[PS12] Pustejovsky, J.; Stubbs, A. “Natural language annotation for machine learning: a guide to corpus-building for applications”. O’Reilly Media, 2012, 342p.

[RG18] Reis, A. C. A.; Gonçalves, V. E. R. “Direito processual penal esquematizado”. Saraiva, 2018, 800p.

[RR02] Reis, E. A.; Reis, I. A. “Análise descritiva de dados”, Relatório Técnico, Departamento de Estatística da Universidade Federal de Minas Gerais, 2002, 36p.

[Say07] Sayão, M. “Verificação e validação em requisitos: processamento da linguagem natural e agentes”, Tese de Doutorado, Pontifícia Universidade Católica do Rio de Janeiro, 2007, 205p.

[SBBM09] Silva, R. C. d.; Bueno, J. L. O.; Bigand, E.; Molin, P. “Multidimensional scaling applied to studies of musical appreciation”, *Paidéia*, vol. 19, Maio-Agosto 2009, pp. 153–158.

[SC14] Souza, E. N. P.; Claro, D. B. “Extração de relações utilizando features diferenciadas para português”, *Linguamática*, vol. 6, Dezembro 2014, pp. 57–65.

[Sou06] Souza, R. R. “Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências”, *Perspectivas em Ciência da Informação*, vol. 11, Maio-Agosto 2006, pp. 161–173.

[Tou18] Tourinho Filho, F. d. C. “Manual de processo penal”. Saraiva, 2018, 320p.

[vBKLK16] van Banerveld, M.; Kechadi, M.-T.; Le-Khac, N.-A. “A natural language processing tool for white collar crime investigation”. In: Transactions on Large Scale Data and Knowledge Centered Systems XXIII, 2016, pp. 1–22.

[VL10] Vieira, R.; Lopes, L. “Processamento de linguagem natural e o tratamento computacional de linguagens científicas”. In: *Linguagens Especializadas em Corpora: Modos de Dizer e Interfaces de Pesquisa*, Editora Universitária da Pontifícia Universidade Católica do Rio Grande do Sul, 2010, pp. 183–201.

[VPV16] Vargas, A. C. G.; Paes, A.; Vasconcelos, C. N. “Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres”. In: Proceedings of the 29th Conference on Graphics, Patterns and Images, 2016, pp. 4.

[Wea55] Weaver, W. “Translation”. In: *Machine Translation of Languages: Fourteen Essays*, Technology Press of the Massachusetts Institute of Technology, 1955, pp. 15–23.

# APÊNDICE A – ALGORITMO PARA TREINO

Algoritmo escrito na linguagem Python utilizado para treino do modelo CNN:

```
from __future__ import unicode_literals, print_function
import plac
import random
from pathlib import Path
import spacy
from spacy.util import minibatch, compounding

def treinamento(model="pt", output_dir="c:/modelo_trio_treinado.md", n_iter=100):
    """Load the model, set up the pipeline and train the entity recognizer.
    """
    if model is not None:
        nlp = spacy.load(model)  # load existing spaCy model
        print("Loaded model '%s'" % model)
    else:
        nlp = spacy.blank("en")  # create blank Language class
        print("Created blank 'en' model")
    # create the built-in pipeline components and add them to the pipeline
    # nlp.create_pipe works for built-ins that are registered with spaCy
    if "ner" not in nlp.pipe_names:
        ner = nlp.create_pipe("ner")
        nlp.add_pipe(ner, last=True)
    # otherwise, get it so we can add labels
    else:
        ner = nlp.get_pipe("ner")
    # add labels
    for _, annotations in TRAIN_DATA:
        for ent in annotations.get("entities"):
            ner.add_label(ent[2])
    # get names of other pipes to disable them during training
    other_pipes = [pipe for pipe in nlp.pipe_names if pipe != "ner"]
    with nlp.disable_pipes(*other_pipes):  # only train NER
        # reset and initialize the weights randomly     but only if we're
        # training a new model
        if model is None:
            nlp.begin_training()
        for itn in range(n_iter):
            random.shuffle(TRAIN_DATA)
            losses = {}
            # batch up the examples using spaCy's minibatch
            batches = minibatch(TRAIN_DATA, size=compounding(4.0, 32.0, 1.001))
            for batch in batches:
```

```
                texts, annotations = zip(*batch)
                nlp.update(
                    texts,  # batch of texts
                    annotations,  # batch of annotations
                    drop=0.5,  # dropout - make it harder to memorise data
                    losses=losses,
                )
            print("Losses", losses)
    # test the trained model
    for text, _ in TRAIN_DATA:
        doc = nlp(text)
        print("Entities", [(ent.text, ent.label_) for ent in doc.ents])
        print("Tokens", [(t.text, t.ent_type_, t.ent_iob) for t in doc])
    # save model to output directory
    if output_dir is not None:
        output_dir = Path(output_dir)
        if not output_dir.exists():
            output_dir.mkdir()
        nlp.to_disk(output_dir)
        print("Saved model to", output_dir)
        # test the saved model
        print("Loading from", output_dir)
        nlp2 = spacy.load(output_dir)
        for text, _ in TRAIN_DATA:
            doc = nlp2(text)
            print("Entities", [(ent.text, ent.label_) for ent in doc.ents])
            print("Tokens", [(t.text, t.ent_type_, t.ent_iob) for t in doc])
treinamento()
```


Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br